Institute of Biotechnology and
Department of Biosciences, Division of Genetics
Faculty of Biological and Environmental Sciences
University of Helsinki
Doctoral Programme in Integrative Life Science (ILS)

and

Department of Clinical Chemistry
Fimlab Laboratories and Finnish Cardiovascular Research Center
Tampere
Faculty of Medicine and Health Technology
Tampere University

Efficient gene set analysis of high-throughput data

From omics to pathway architecture of health and disease

Pashupati Prasad Mishra

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biological and Environmental Sciences of the University of Helsinki, for public examination in Room 4, Metsätalo, on 6th of August, 2020, at 12 o'clock.

Helsinki 2020

Supervised by
Professor Liisa Holm
Faculty of Biological and Environmental Sciences & Institute of Biotechnology
University of Helsinki, Finland

Dr. Petri Törönen
Institute of Biotechnology
University of Helsinki, Finland

&

Professor Terho Lehtimäki
Department of Clinical Chemistry
Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere
Faculty of Medicine and Health Technology
Tampere University, Finland

Reviewed by
Professor Sangita Kulathinal
Department of Mathematics and Statistics
University of Helsinki, Finland

Assistant Professor Juulia Jylhävä
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Sweden

Opponent
Dr. Alvis Brazma
European Bioinformatics Institute (EMBL-EBI)
United Kingdom

Custos
Professor Liisa Holm
Department of Biolocal and Environmental Sciences & Institute of Biotechnology
University of Helsinki, Finland

ISBN 978-951-51-6287-8 (paperback)
ISBN 978-951-51-6288-5 (PDF)

https://ethesis.helsinki.fi
Unigrafia Oy, Helsinki 2020

Contents

List of original publications

This thesis is based on the following articles, which are referred to in the text by their Roman numerals.

I. **Mishra, PP.**, Törönen, P., Leino, Y., & Holm, L. (2014). **Gene set analysis: limitations in popular existing methods and proposed improvements.** *Bioinformatics*, 30(19), 2747-2756.

II. **Mishra, PP.**, Medlar, A., Holm, L., & Törönen, P. (2016). **Robust multi-group gene set analysis with few replicates.** *BMC Bioinformatics*, 17(1), 526.

III. **Mishra, PP.**, Hänninen, I., Raitoharju, E., Marttila, S., Mishra, BH., Mononen, N., Kähönen, M., Hurme, M., Raitakari, O., Törönen, P., Holm, L. & Lehtimäki, T. (2020). **Epigenome-450K-wide methylation signatures of active cigarette smoking: the Young Finns Study.** *In press: Accepted for publication in Bioscience Reports.*

The publications were adapted with the permission of the copyright owners.

Abstract

Background A wide range of diseases, normal variations in physiology and development of different species are caused by alterations in gene regulation. The study of gene expression is thus crucial for understanding both normal physiology and disease mechanisms. High-throughput measurement technologies allow the profiling of tens of thousands of genes simultaneously. However, the high volume of data thus generated poses methodological challenges in inferring biological consequences from gene expression changes. Traditional gene wise analysis of high dimensional data is overwhelming, prone to noise and unintuitive. The analysis of sets of genes (gene set analysis, GSA), solves the problem by boosting statistical power and biological interpretability. Despite more than a decade of research on gene set analysis, there are still serious limitations in the existing methods.

Aims of the study The objectives of this study were: (1) development of an efficient p-value estimation method for GSA; (2) development of an advanced permutation method for GSA of multi-group gene expression data with fewer replicates; and (3) implementation of the developed methods for the identification of novel smoking induced epigenetic signatures at biological pathway level.

Materials and methods The first study involved the assessment of four different statistical null models for modeling the distribution of gene set scores calculated with the Gene Set Z-score (GSZ) function from permuted gene expression data. A new GSA method - modified GSZ (mGSZ) - based on GSZ and the most optimal distribution model was developed. mGSZ was evaluated by comparing its results with seven other popular GSA methods using four different publicly available gene expression datasets. The second study involved the evaluation of six different permutation schemes for GSA of multi-group (more than two groups) datasets based on the identification of reference gene sets generated using a novel data splitting approach. A new GSA method based on a modification of mGSZ (mGSZm) was de-

veloped by implementing the best permutation method for the analysis of multi-group data with fewer than six replicates per group. mGSZm was evaluated by contrasting its performance with seven other state-of-the-art GSA methods suitable for multi-group data. The evaluation was based on three different publicly available multi-group datasets. The third study involved an implementation of mGSZ for GSA of genome-wide DNA methylation data from the Cardiovascular Risk in Young Finns study (YFS) cohort with gene sets downloaded from the Molecular Signature Database (MSigDB). Methylation measurements were done on a subset of 192 individuals from whole-blood samples from the 2011 follow-up study using Illumina Infinium HumanMethylation450 BeadChips.

Results Overall, efficient and robust GSA methods were developed (studies I-II) and implemented (study III). In study I, the results demonstrated a clear advantage of asymptotic p-value estimation over empirical methods. mGSZ, a GSA method based on asymptotic p-values, requires fewer permutations which speeds up the analysis process. mGSZ outperformed state-of-the-art methods based on three different evaluations with three different datasets. In study II, results from a novel evaluation approach with two different datasets suggested that the proposed advanced permutation method outperformed the naive permutation method in GSA of multi-group data with fewer than six replicates. Evaluation of mGSZm, a GSA method equipped with the advanced permutation method and asymptotic p-value estimation method, showed that the method is robust and, despite only using fewer than six replicates, is able to consistently identify a high proportion of relevant gene sets in three different multi-group datasets. In study III, GSA of YFS methylation data using mGSZ identified a total of 13 significant (false discovery rate ≤ 0.05) smoking related pathways. The results included a novel finding that the highly regenerating olfactory sensing system responds to tobacco smoke and toxin exposure through epigenetic mechanisms. Besides the novel finding, the study also confirmed previous findings of smoking induced alteration in methylation in biological pathways involved in lung tissue repair and maintenance, chronic inflammation, lung cancer, platelet function, thrombosis and nervous system development.

Conclusions In this thesis: i) a novel approach of efficient p-value estimation for permutation based GSA was introduced, ii) an advanced permutation method for GSA of multi-group data with fewer than six replicates was developed, and iii) a novel smoking related epigenetic signature was

identified using the developed methods.

Abbreviations

CAMERA . . Correlation Adjusted MEan RAnk gene set test
EVD . . . . . Extreme Value Distribution
FDR . . . . . False Discovery Rate
GAGE . . . . . Generally Applicable Gene-set Enrichment
GEVD . . . . . Generalized Extreme Value Distribution
GO . . . . . Gene Ontology
GSA . . . . . Gene Set Analysis
GSEA . . . . . Gene Set Enrichment Analysis
GSZ . . . . . Gene Set Z-score
KEGG . . . . . Kyoto Encyclopedia of Genes and Genomes
KS . . . . . Kolmogorov Smirnov test
mAllez . . . . . Modified Allez
mGSA . . . . . Modified Gene Set Analysis
mGSZ . . . . . Modified Gene Set Z-score
mGSZm . . . . Modified Gene Set Z-score for Multi-group data
MSE . . . . . Mean Squared Error
MSigDB . . . . Molecular Signature DataBase
ORA . . . . . Over Representation Analysis
QuSAGE . . . Quantitative Set Analysis for Gene Expression
ROAST . . . . Rotation Gene Set Testing
TF . . . . . Transcription Factor
wKS . . . . . Weighted Kolmogorov Smirnov test
WRS . . . . . Wilcoxon Rank-Sum test
YFS . . . . . Young Finns Study

Chapter 1

Introduction

1.1 Transcriptomics and epigenetics

1.1.1 Central Dogma of Molecular Biology

The Central Dogma of Molecular Biology describes the process of gene expression. Gene expression is the process by which the genetic code in the DNA (deoxyribonucleic acid) is used for the synthesis of functional gene products like proteins and RNAs. DNA is the hereditary material in humans and most other organisms. Most DNA is located in cell nucleus densely packed into thread-like structures called chromosomes. The part of DNA that is present in mitochondria is called mitochondrial DNA. DNA is made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The sequence of these chemical bases, about 3 billion in total in humans, form the genetic code of DNA [22]. Genes, the basic physical and functional unit of life, are segments of the DNA string that are transcribed into RNA and then translated into a functional protein. In humans, there are estimated to be 20000 to 25000 protein-coding genes with size varying from a few hundred to more than 2 million bases [105]. Genes carry the instruction needed to code a protein.

1.1.2 Transcriptomics

In gene expression, the first step is transcription that involves the formation of protein-coding or non-coding regulatory RNA molecules from a gene with the help of an enzyme called RNA polymerase. The collection of RNA molecules transcribed in a cell and a tissue is called the transcriptome. The second step is translation when messenger RNA (mRNA) is decoded to pro-

duce a specific amino acid chain with the help of ribosomes. The amino acid chain folds into an active protein and performs its functions in the cell. Gene expression, thus, acts as a proxy for translation and translation events that represent a snapshot of a biological state.

Assuming that a normal cell has a standard gene expression profile, a shift in the profile is assumed to be indicative of altered protein level and consequently an altered biological state. Thus, gene expression measurement and analysis is important for studying the molecular mechanism of an altered biological state like disease, the identification of diagnostic or prognostic markers, the classification of diseases, monitoring the response to a therapy and understanding the molecular mechanism of biological processes. However, it is important to note that several studies have reported poor agreement between mRNA levels and the corresponding protein levels [76, 71] which could be due to several reasons such as complicated post-transcriptional mechanisms, different in vivo half-lives of proteins, error and/or noise in protein and mRNA measurements and different protein turnover rates.

Advances in molecular biology technologies allow genome-wide profiling of many genes simultaneously. Analysis of such high-throughput gene expression data is used to identify genome-wide differences in the levels of gene transcription between different experimental groups such as healthy controls and people with certain disease. DNA microarrays, invented in the 1990s, are microscope slides with thousands of tiny spots at defined positions [72]. The spots contain DNA molecules that act as probes to detect gene expression. Probes are specific DNA molecule designed to target a gene or other DNA element by DNA hybridization with a complementary DNA strand. The probe-target gene hybridization generates fluorescence, the intensity of which is detected and quantified to estimate the relative expression level of target genes. The underlying rationale is that gene expression levels are directly proportional to the level of probe-gene hybridization. Gene expression analysis starts with experimental design that starts with sample collection. During sample collection, care should be taken in minimizing sources of variation other than those under investigation. For example, in a case-control study, matching or pairing of the samples must be done appropriately to avoid the effects of a confounding variable. The decision of the number of technical and/or biological replicates is another aspect of experimental design for reliable statistical inference. Generally, three replicates are a minimum requirement for a reliable statistical test.

Once the sample collection is done, mRNA samples (for example, from whole blood) are collected, labelled and applied to the microarray. Proper quality assurance should be taken during collection, isolation, storage and preservation of RNA/DNA samples. A sufficient amount of high quality or intact RNA is crucial in any molecular biology experiments. Analysis of the generated data must account for technical artifacts that can arise from sources such as different date, lab, technician, microarray chip or failed hybridization.

Despite their popularity, microarrays have several limitations. Measurement accuracy for genes that are mildly expressed is limited by background hybridization. An experiment is limited to the genes which have probes available in the microarray. A more recent high-throughput sequencing technology called as RNA sequencing (RNA-Seq) addresses the limitations by directly sequencing the transcripts [86]. RNA-seq has rapidly replaced microarrays for whole-genome transcriptome profiling because of its ability to detect novel transcripts, allele-specific expression and splice junctions [112]. Recently, there has been growing interest in transcriptomics of single cells through single-cell RNA sequencing technologies. This holds great potential in revealing biological insights of health and disease at cellular level as compared to traditional bulk population based methods [115].

1.1.3 Epigenetics

Epigenetics is the study of heritable changes in gene expression that is not due to any changes of the DNA sequence itself. Virtually all cells in an organism contain the same DNA. Yet, not all genes are expressed simultaneously in all cell types. Epigenetic changes determine whether or not a particular gene is expressed in a particular cell type, thereby influencing the production of cell specific proteins. For example, genes that code for protein needed for bone growth are turned on in bone cells but turned off in muscle cells. The epigenome constitutes all the chemical compounds that regulate the activity (expression) of all the genes within the genome. The chemical compounds are not part of the DNA sequence but are attached to DNA and can be inherited through generations. The epigenome can be affected by environmental factors like diet, smoking and pollutants. Three major epigenetic changes are DNA methylation, histone modification and non-coding RNAs (Figure 1.1) [94]. Histones are proteins responsible for condensing the DNA of eukaryotic cell nuclei and organizing it into units called nucleosomes. Post-translational modification to histone proteins such as

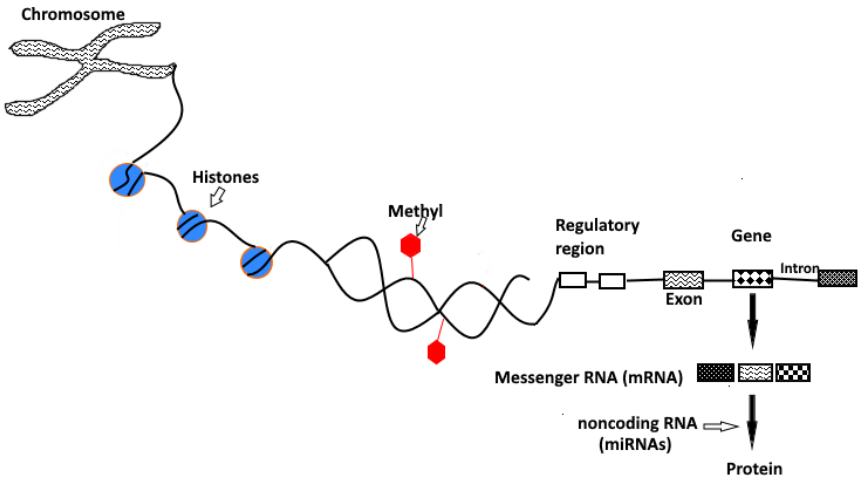


Figure 1.1: A schematic diagram showing different epigenetic mechanisms; histone modification, methylation and posttranscriptional regulation by MicroRNAs (miRNAs).

methylation, phosphorylation, acetylation, ubiquitylation and sumoylation changes chromatin structure and consequently alters gene expression. Noncoding RNAs (ncRNAs) are RNA molecules that are transcribed from DNA but not translated into proteins [48]. These molecules can be categorized into two major groups; the short ncRNAs (<30 nucleotides) and the long ncRNAs (>200 nucleotides). Short ncRNAs include microRNA (miRNA), short interfering RNA (siRNA) and piwi-interacting RNA (piRNA). MicroRNAs, which are small RNA molecules of 17 to 24 nucleotides, are one of the most widely studied non-coding RNAs. MicroRNA silences gene expression by binding to the 3′UTR of its target-gene's mRNA.

DNA methylation is one of the major epigenetic factors influencing gene expression through the addition of methyl groups to the DNA molecule. DNA bases (cytosine and adenine) can be methylated. Adenine methylation has been observed in bacterial, plant and recently also in mammalian DNA [106]. Cytosine methylation which is common in bacteria and eukaryotes,

is the most studied type of methylation. Cytosine methylation involves the transfer of a methyl group from S-adenyl methionine to the fifth carbon of cytosine residue catalyzed by a family of DNA methyltransferases to form 5-methylcytosine. The process plays a critical role in the regulation of gene expression and consequently in mammalian development and diseases [44].

Genome-wide profiling of human DNA methylation used in this study was based on whole blood sample and Infinium HumanMethylation450K BeadChip (450K methylation array). The high-throughput technology is based on the following principle. Bisulphite treatment of DNA converts cytosine residues to uracil. However, methylated cytosine (5-methylcytosine) residues are unaffected by this treatment. This property of cytosine residues provides the "gold standard" approach to assess the DNA methylation status at single-nucleotide resolution. CpG sites are regions on DNA where cytosine and guanine appear consecutively in the 5′to 3′direction. The "p" stands for the phosphodiester bond that joins the two nucleotides. CNG sites, on the other hand, are sites with consecutive cytosine and guanine nucleotides with any nucleotide in between. The array includes a total of 485,764 cytosine sites out of which 482,421 are CpG (99.3%) and 3,343 are CNG (0.7 %) sites. It is based on two different chemistries; Infinium I and Infinium II. Infinium I contains two probes per methylation site, one for methylated and another for unmethylated query site. The 3′terminus of the probes is designed such that it matches either the methylated cytosine or the thymine base (resulting from bisulphite treatment or present in the genome). Infinium II, on the other hand, has only one probe per methylation site. The 3′terminus of the Infinium II probe is designed such that it complements the base directly upstream of the query site. Identification of methylation is based on single base extension that results in addition of a labeled G or A base, complementary to either the "methylated" C or "unmethylated" T.

1.2 Gene Set Analysis

1.2.1 Motivation

In recent years, the advent of high-throughput molecular profiling techniques and their rapid evolution thereafter have revolutionized the way bioscience research is done. For example, there has been a major shift in the approach to studying disease, from the traditional individual molecule-wise study to system-level studies. With this benefit, there also came a

challenge to be able to interpret the huge list of molecules generated from the analyses of high-throughput data, which is often noisy.

The fundamental goal of genome-wide omics experiments is the detection of biological processes or pathways that behave consistently differently between groups of samples under different conditions. The traditional approach for analyzing such data is the identification of individual differentially expressed molecules such as genes. A typical approach is to perform statistical tests such as analysis of variance (ANOVA) and select a list of genes based on an arbitrary p-value threshold. The selected genes are studied individually for their biological relevance. Gene-wise analysis of genome-wide data for that purpose has the following problems: i) no single gene may be differentially expressed based on the chosen threshold, ii) an overwhelming number of genes may be differentially expressed without a unifying biological theme (unintuitive), iii) differentially expressed genes with multiple functions are hard to interpret.

One approach to address the problems is to analyze the list of molecules group-wise, where the grouping is based on some prior knowledge. The idea was introduced by [117] who analyzed microarray data by first grouping open reading frames into clusters based on the similarity of their expression patterns, followed by enrichment analysis of the clusters against pre-defined functional categories. The approach was originally developed of analyzing a list of genes and hence called as GSA. The approach solves the problems as: i) it increases the statistical power as smaller but consistent changes in all the genes in a set is likely to stand out clearly as opposed to individual genes (larger signal to noise ratio), ii) gene level noise signals cancel out when analyzed as a group, iii) it provides intuitive results if the genes in a set are related with respect to a common biological theme. The approach involves the calculation of gene set scores analogous to gene scores such as t-scores or fold change. The approach has several names such as “gene set analysis”, “functional annotation”, “pathway analysis”, “gene set enrichment analysis” and “gene list enrichment analysis”. From here onwards, we refer the approach collectively as “Gene Set Analysis (GSA)”. Even though the earlier gene set analysis methods were developed particularly for transcriptomics and differential expression analysis, the approach is applicable to all types of genome-wide omics data ([100]).

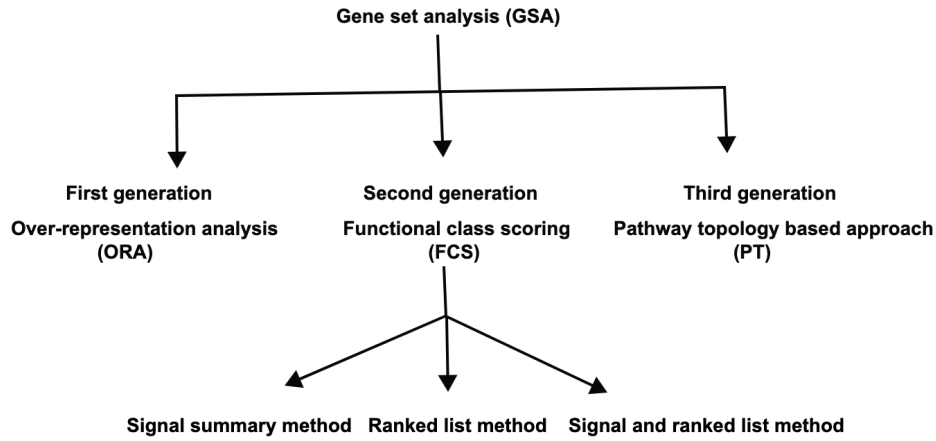


Figure 1.2: A schematic diagram showing different types of gene set analysis.

1.2.2 Gene sets

Gene sets are genes grouped together based on shared biological features like biological pathways, transcription factors or chromosomal location. The type of gene sets to be used in an analysis depends on the research question. Several databases or resources such as Gene Ontology (GO) [5], [24], Kyoto Encyclopedia of Genes and Genomes (KEGG) [56], Reactome Pathway Database [28] and Molecular Signatures Database (MSigDB) [114], [68] provide different types of gene set collections.

1.2.3 Types

Many different GSA methods have been developed over the last decade. The methods can be categorized into different types described in this section (Figure 1.2.3).

GSA methods can broadly be categorized into **three generations**. Methods belonging to the **first generation** are the earliest GSA methods commonly called over-representation analysis (ORA) or enrichment analysis. Over-representation based methods determine the overlap between a test gene list and a curated database like gene ontology, looking for overlaps that are bigger than that expected by chance. The most common method to obtain the list of genes is to get a list of differentially expressed genes

based on an arbitrary threshold such as p-value and/or fold change. The idea was first implemented by [117]. ORA methods are based on statistical tests like the hypergeometric test, Fisher's exact and chi-squared (i.e the problem can be formulated in point of view of different statistical tests). There are several freely available tools for ORA [11], [50]. Some of the ORA methods take into account the differential expression as a score calculated as a product of p-value and fold change. These methods are quick and easy to implement but have several limitations, including:

(i) Applicable only when differentially expressed genes are found

(ii) Results change with arbitrary p-value thresholds

(iii) Disregards gene-gene correlation

(iv) Massive reduction in sensitivity

Second generation GSA methods are called functional class scoring (FCS) methods. FCS methods are threshold free (i.e., they do not require p-value thresholds) and thus utilize all the genes in a gene set (whether or not differentially expressed) to calculate a gene set score. In contrast to ORA, threshold based selection of gene list is not required, which guarantees a unique result for each dataset. Furthermore, this approach allows the identification of significant gene sets among weakly but coherently regulated genes. These methods can be based on signal summary scores [90, 118] or ranked lists [84, 12, 17]. Signal summary based methods use the whole gene list and calculate a gene set score as a function of the gene set members' gene scores (e.g., t-scores or fold change scores). An obvious drawback of the signal summary approach is that the method assumes homogeneous behavior of the members of a gene set. This assumption is violated in cases where a gene set contains both up and down-regulated genes or where genes have been misclassified as members of the gene set. Ranked list based methods go through the whole ordered gene list and analyze over-representation of the gene set at every possible threshold and the threshold with the highest score is selected. A drawback with this approach is that it ignores scores associated with genes (example, t-scores or fold change). The assumption of a constant step between consecutive genes is violated in real datasets as there are usually drastic changes in gene expression at the tails of ordered gene lists. [114] proposed a GSA method combining signal summary and rank list. The combination mitigates the above mentioned limitations of signal summary and ranked list based methods. The method calculates gene set statistics based on over– or under-representation and gene level

scores over the ordered gene list to find the strongest signal.

Third generation GSA methods, called the pathway topology-based approach, incorporate biological network structure into the analysis to account for gene-gene correlation [83]. These methods take gene-gene correlation into account by weighting genes based on their relationship with other genes in a gene set. For example, a gene that is correlated to only one other gene in a gene set is given less weight as compared to a gene that is correlated to more than one gene in the gene set. In addition to gene sets data, these methods require gene-gene interaction data or knowledge of pathway structure.

GSA categorization based on Null hypothesis

Based on the null hypothesis tested, GSA methods can be either competitive or self-contained. The null hypothesis of self-contained methods states that none of the genes in a gene set are associated with the phenotype. These methods are useful when a researcher is interested in testing whether one or a few gene sets of interest is/are associated with the phenotype. Methods proposed by [42], [77], [60], [30], and [126] are examples of self-contained gene set analysis methods. Competitive gene set analysis methods test whether the genes in a gene set are more associated to the phenotype as compared to other genes (genes not in the analyzed gene set). These methods are used to rank collection of gene sets based on the strength of their association with the phenotype. GSEA [114], GSA [34] and CAMERA [127] are examples of competitive gene set analysis.

This thesis is focused on signal and ranked list based functional class scoring GSA methods that test competitive null hypothesis. For simplicity, we will refer this class of methods as gene set analysis (GSA) methods in the rest of the thesis.

1.2.4 Gene set scoring functions

GSA involves two parts; 1) calculation of a gene set score as a function of gene level statistics (for example t-scores or fold change scores) of member genes of the analyzed gene set, and 2) estimation of the statistical significance of the gene set score. The gene set score quantifies the magnitude by which genes in a gene set are overrepresented in either end of the ranked gene list. This Section describes popular gene set scoring functions used in

this thesis.

Kolmogorov-Smirnov statistic (KS-statistics)

The Kolmogorov-Smirnov test is a non-parametric test, i.e., the test does not require data to follow any known statistical distribution distribution. The purpose is to either test for differences in the shape of two sample distributions or to compare the distribution of a sample to an expected statistical distribution. The test compares the overall shape of distributions instead of central tendency, dispersion or other parameters. Thus, the null hypothesis for two sample test is that two samples are drawn from populations with the same distribution. The KS-statistics is simply the maximum absolute difference between the two cumulative distribution functions. The gene set score in GSEA [114] is based on KS-statistics which is calculated as follows;

1. Perform differential gene expression analysis and rank the genes based on their differential gene expression scores, for example t-scores.

2. Compute cumulative sum over the ranked genes as follows:

   (a) Increase sum if gene is in the analyzed gene set, decrease it otherwise.

   (b) The magnitude of the increment depends on the strength of association of the gene with phenotype.

3. The maximum deviation from zero is the gene set score.

4. Normalization of Enrichment Score for multiple testing correction step.

The maxmean statistic

[34] proposed maxmean statistics as a gene set score and showed that this statistic is more powerful than the modified KolmogorovSmirnov statistic used in GSEA. For a given gene set, the mean of the positive or negative part of gene scores (for example, differential expression test scores) is calculated. The maxmean statistics is the value that is larger in absolute value. Note that the mean of the positive or negative part of gene scores is calculated by dividing the sum by total number of genes in the gene set instead of total number of positive or negative gene scores. This makes the gene set score sensitive to large gene scores in either or both directions.

Wilcoxon rank-sum statistic

The Wilcoxon Rank-Sum (WRS) test is a rank based non-parametric test for comparing two groups of observations without the assumption of any distribution. The gene set score is simply the sum of ranks of member genes of the analyzed gene set in the whole gene list. [122] proposed a two-sided WRS test to compare the ranks of the genes belonging to a gene set with the ranks of the remainder of the genes. WRS is also implemented in SAFE (Significance Analysis of Function and Expression) [10] as the gene set scoring function.

Random-set enrichment score (Allez)

The random-set enrichment score is based on the comparison of the sum of differential expression scores of gene set members to the expected average of a random set of genes of the same size and also to the expected variance of a random set of genes of the same size [88]. For the analyzed gene set, the score is first calculated as a raw average of gene scores in the set. The score is a random variable because the gene set is a random set of genes among the gene pool in the data. The enrichment score is based on the comparison of the analyzed gene set score with a hypothetical gene set from the same dataset. As the analyzed gene set can be considered as a random gene set with m genes drawn uniformly from the pool of genes in the dataset without replacement, it is equivalent to gene label permutation. The gene set score is obtained by standardization of the raw score (average of member gene scores of the set) by substracting the global mean of gene scores and dividing the difference with global variance that depends on the size m of the gene set. The method has been implemented in an R package *Allez*. For simplicity, we refer this method as *Allez* in the rest of the text.

Sum (SUM) and sum of squares (SS) based GSA scores

Gene set scoring functions based on the sum and sum of squares were also used as reference GSA methods in this thesis. The sum statistic is the simple sum of the differential gene expression test scores of the member genes of the analyzed gene set with background subtraction. The sum of squares statistic is simply the squared version of SUM.

Gene Set Z-score (GSZ)

GSZ combines the signal summary and ranked list based approach by calculating a statistic that is based on under– or over-representation and by

analyzing the whole gene list for the strongest signal. The approach is similar to that of GSEA except that GSZ involves the derivation of the mean and standard deviation for normalizing the gene set score as a Z-score.

Input The first step in gene set analysis with GSZ involves the calculation of gene scores. Gene scores can be: i) differences in the gene expression profiles between two compared groups such as treatment and control groups, or ii) the association of the gene expression profile with a continuous phenotype such as waist size measurement of a cohort participants. The genes are then sorted based on the gene scores and used as input for gene set analysis.

Derivation Given a ranked gene list, a subset is taken for gene set score calculation by placing a threshold. The process of selecting subsets of the ranked list and calculating the score is done repetitively by placing the threshold consecutively after each gene. Thus, the analysis involves the calculation of scores for subsets ranging in size from one gene to all the genes in the ranked list. The following list of notations will be used in the derivation;

- M: Total number of genes in a subset
- $X_i, i = 1, ..., M$: Gene score for ith gene.
- N: Number of genes that belong to analyzed gene set (member genes)
- L: Total number of genes in the whole gene list
- K: Number of member genes in the whole gene list
- S: Sum of expression levels of N genes

For a given subset from a ranked gene list, the gene set score is simply a difference between the sums of member and non-member genes,

$$Diff = \sum X_{member} - \sum X_{non-member}, \tag{1.1}$$

which is calculated separately for each subset and analogously for the lower part of the ranked gene list. The calculated difference is unstable due to its

sensitivity to differences in size of analyzed gene sets, ranked gene subsets and variances of the gene scores in different subsets. The instability can be solved by normalizing the difference (equation 1.1) with its estimates of expected value and variance under the null hypothesis that the member and non-member genes are distributed randomly across the ranked gene list,

$$Z = \frac{Diff - E(Diff)}{\sqrt{D^2(Diff)}} \tag{1.2}$$

where $E(Diff)$ is the expected value of the difference given by,

$$2E(X)E(N) - ME(X),$$

and $D^2(Diff)$ is the variance of the difference given by,

$$4(\tfrac{D^2(X)}{M-1}(E(N)(M - E(N)) - D^2(N)) + E(X)^2 D^2(N)),$$

where $E(N)$ is the mean and $D2(N)$ is the variance of the hypergeometric distribution of the number of genes, N, in the analyzed gene set for the analyzed subset. $D2(X)$ and $E(X)$ are the variance and the mean of the differential expression test scores for the subset.

Derivation of expected value The expected values for the sum of member or non-member gene scores in the equation 1.1 can be defined using conditional probabilities:

$$\begin{aligned} E(S|M, L, K) &= E(\sum X_n) \\ &= \sum P(N = i)E(S|N = i), \end{aligned} \tag{1.3}$$

which is the sum of expected value of S conditional on the number of member genes weighted with the probability of having that number of member genes in the analyzed subset. Equation 1.3 is equivalent to:

$$E(S|M,L,K) = \sum P(N=i)iE(X), \tag{1.4}$$

as the expected value of the sum conditional on the number of member genes is equivalent to the expected value of gene scores of member genes of the analyzed subset multiplied by the number of member genes. The expected value of equation 1.1 can be defined as:

$$\begin{aligned} E(Diff|M,L,K) &= \sum_i P(N=i)iE(X) - \sum_i P(N=i)(M-i)E(X) \\ &= E(X)(\sum_i P(N=i)i - \sum_i P(N=i)(M-i)) \\ &= E(X)(2\sum_i P(N=i)i - M) \\ &= E(X)(2E(N) - M) \\ &= 2E(X)E(N) - ME(X), \end{aligned} \tag{1.5}$$

which leads to a simple function of: i) the expected value of hypergeometry distribution $E(N)$, ii) the expected value of gene scores in the analyzed subset $E(X)$, and iii) the number of genes M in the analyzed subset. $E(X)$ is calculated by taking empirical mean from the whole subset.

Derivation of variance The variance for sum of member or non-member gene scores in the equation 1.1 can be defined as:

$$\begin{aligned} D^2(S|M,L,K) &= E(S^2) - E(S)^2 \\ &= \sum_i P(N=i)E(S^2|N) - E(S)^2, \end{aligned} \tag{1.6}$$

For simplicity, lets consider the case where N is fixed. Then, the expectation $E(S^2|N)$ can be expressed as a sum of the variance and squared expectation,

$$\begin{aligned} E(S^2|N) &= D^2(S|N) + E(S|N)^2 \\ &= D^2(X_i)\frac{N(M-N)}{M-1} + E(X)^2N^2, \end{aligned} \tag{1.7}$$

The first term of the equation 1.7 represents the variance of the sum of N values selected from a pool of X_1, X_2, $..X_M$ values with variance $D^2(X)$ which is estimated empirically from the subset. The derivation of $D^2(S|N)$ is taken from supplementary text S1 of the original GSZ article [119]. Substituting equation 1.9 to equation 1.6 gives the following:

$$\begin{aligned} D^2(S|M,L,K) &= \sum_i P(N=i)E(S^2|N) - E(S)^2 \\ &= \sum_i P(N=i)(D^2(X_i)\frac{i(M-N)}{M-1} + E(X)^2i^2) \\ &- E(X)^2E(N)^2 \\ &= D^2(X)\sum_i P(N=i)\frac{(Mi-i^2)}{M-1} + E(X^2)\sum_i P(N=i)i^2 \\ &- E(X)^2E(N)^2 \\ &= \frac{MD^2(X)}{M-1}\sum_i P(N=i)i - \frac{D^2(X)}{M-1}\sum_i P(N=i)i^2 \\ &+ E(X)^2(E(N^2) - E(N)^2) \\ &= \frac{MD^2(X)E(N)}{M-1} - \frac{D^2(X)(D^2(N)+E(N)^2)}{M-1} + E(X)^2D^2(N) \\ &= \frac{D^2(X)}{M-1}(E(N)(M-E(N)) - D^2(N)) + E(X)^2D^2(N) \end{aligned} \tag{1.8}$$

The final derivation of equation 1.8 involves equations of the mean and variance of the hypergeometric distribution ($E(N)$, $D^2(N)$), the gene score of the subset of data ($E(X)$, $D^2(X)$) and the size of the subset (M). Equation 1.8 represents the variance of one of the summations of the equation 1.1. The complete variance in equation 1.1 can be derived by multiplying equation 1.8 with the squared constant 2^2. The final Gene Set Z-score is thus obtained as:

$$\begin{aligned}
Z &= \frac{Diff - E(Diff)}{\sqrt{D^2(Diff)}} \\
&= \frac{Diff - E(X)(2E(N) - M)}{2\sqrt{D^2(S)}} \\
&= \frac{\sum X_{member} - \sum X_{non-member} - E(X)(2E(N) - M)}{2\sqrt{D^2(S)}} \\
&= \frac{\sum X_{member} - ME(X) + \sum X_{member} - E(X)(2E(N) - M)}{2\sqrt{D^2(S)}} \\
&= \frac{2\sum X_{member} - ME(X) - E(X)(2E(N) - M)}{2\sqrt{D^2(S)}} \\
&= \frac{2\sum X_{member} - ME(X) - 2E(X)E(N) + ME(X)}{2\sqrt{D^2(S)}} \\
&= \frac{2\sum X_{member} - 2E(X)E(N)}{2\sqrt{D^2(S)}} \\
&= \frac{\sum X_{member} - E(X)E(N)}{\sqrt{D^2(S)}}
\end{aligned} \tag{1.9}$$

1.2.5 P-value estimation

The P-value is the probability of seeing the observed or more extreme results when the null hypothesis is true. P-values and statistical significance have been recently criticized [2]. However, p-values were used in this thesis work as a way to estimate the most informative gene sets for biological analysis. Based on the method used to estimate p-value, GSA can be either permutation based or parametric. Parametric p-values are calculated from an assumed underlying distribution of gene set scores. For example, Irizarry et al., [51] and Kim and Volsky [58] proposed fully parametric gene set analysis methods based on a normal approximation of the gene set scores. Strong statistical assumptions of the parametric approach are not

always met. Permutation based p-value can be calculated by permuting either genes or samples. Sample-wise permutation preserves the correlation structure among genes in a gene set and is thus more preferable. Ideally, p-value estimation is done by comparing the original test statistics with test statistics obtained from all possible permutations of the sample labels. However, it is computationally not feasible to consider all possible permutations to generate the null distribution. For example, in case of a moderately big gene expression dataset with two groups and 10 biological replicates per group, The total number of permutations while excluding mirror image is $\frac{1}{2}\frac{(20)!}{10!(20-10)!} = 92378$. Generally, in practice a subset of permutations are chosen for p-value estimation, the most common in GSA being 1000-2000 permutations (Figure 1.3). The problem with this approach is that the minimum obtainable p-value varies with the chosen number of permutations. Gene set scores with real p-values less that 1/(number of permutations) will be assigned zero thus limiting researcher's ability to rank the gene sets based on their significance.

1.2.6 Permutation

A permutation test is a non-parametric statistical procedure to determine the statistical significance of a test statistic based on rearrangements of the class labels of a dataset. The test essentially constructs a null distribution (sampling distribution for the test statistic under the null hypothesis) instead of assuming one as is the case with parametric methods. Permutation tests are useful in situations where there are insufficient data or evidence to support a particular statistical model for the analyzed measurements.

The null hypothesis, in the context of GSA, is tested by first calculating the gene set score, the test statistic with the original gene expression data. Next, the null distribution is computed with a permutation test. A large number of samples (gene set scores) under the null hypothesis is needed to estimate the sampling distribution. A large number of datasets can be generated by random rearrangement of the sample labels (referred to as permutations hereafter). A gene set score is calculated from each of the rearranged or permuted data. Permuting of sample labels will have no effect on the outcome when null hypothesis is true. The P-value for the gene set score is estimated by ranking the original gene set score among the permuted gene set scores.

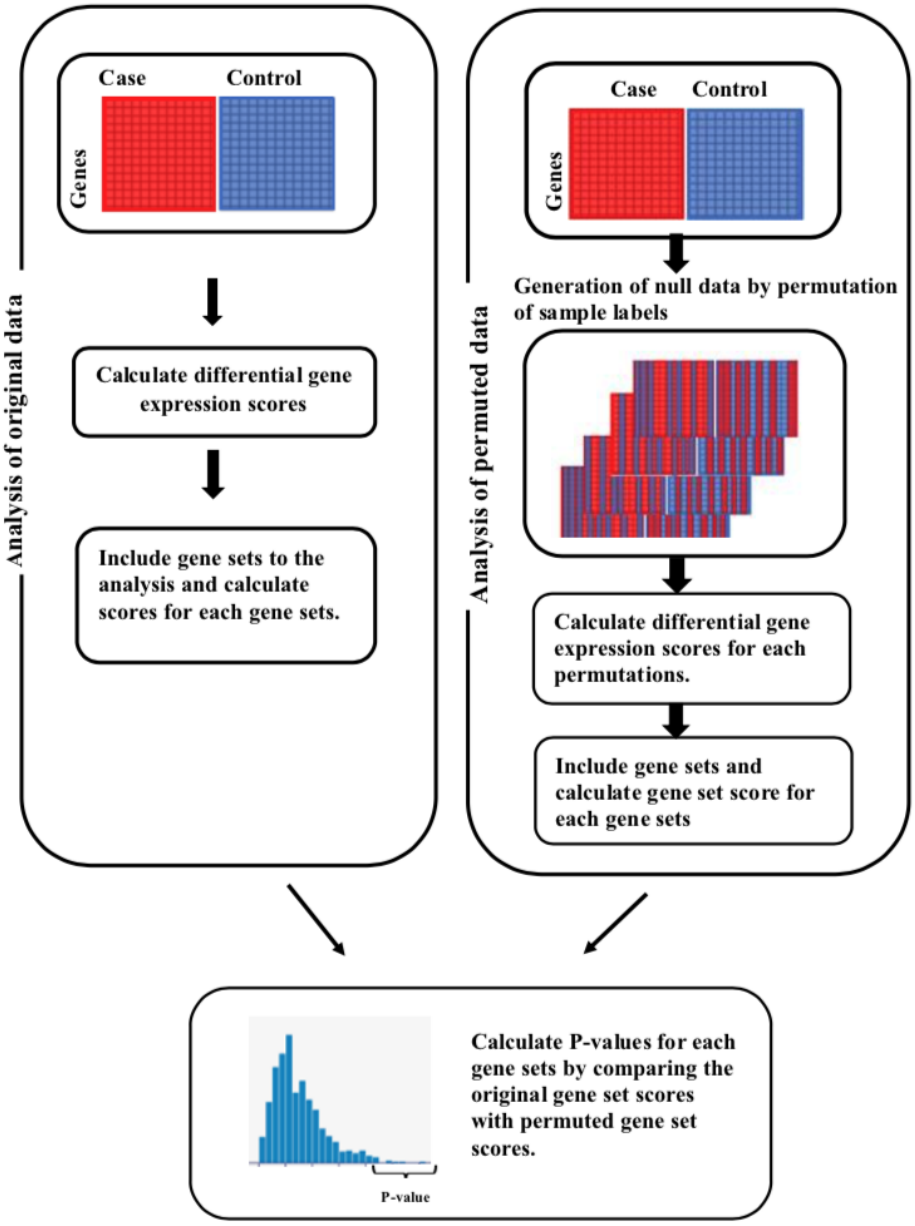


Figure 1.3: Diagrammatic representation of sample permutation based competitive gene set analysis workflow.

Chapter 2

Aims of the present study

The aims of this study were: (a) to develop novel sample permutation based competitive gene set analysis methods addressing the limitations in the most popular current methods, and (b) to implement the developed methods to identify active smoking induced epigenetic signatures in biological processes. The specific aims were covered in the following three publications:

I. Develop efficient p-value estimation method for permutation based gene set analysis (Publication I).

II. Develop advanced permutation method for sample permutation based gene set analysis of multi-group data with fewer than six replicates (Publication II)

III. Identify smoking related epigenetic signatures by gene set analysis of genome-wide methylation data (Publication III).

Chapter 3

Materials and Methods

3.1 Asymptotic p-value estimation

P-value estimation by fitting several candidate statistical distribution models (Section 3.1.1) on a null distribution of gene set scores generated from permuted gene expression data was evaluated. The most optimal distribution model based on the evaluations was applied to GSZ and the updated method is called modified GSZ (mGSZ). mGSZ was evaluated by comparing its performance with seven other popular gene set analysis methods (Section 3.1.3) on multiple datasets (Section 3.1.4).

3.1.1 Extreme Value Distribution

Extreme value distributions are used to model extreme or rare events, for example, extreme flood, temperature or snowfall. Extreme value theory that deals with such events states that the maximum or minimum of the collection of random observations from the same distribution can be modeled with the extreme value distribution [61]. The class of extreme value distribution involves three types of distributions - type I (Gumbell), II (Frechet) and III (Weibull).

The GSZ score of a gene set represents the absolute maximum score among the scores calculated over all the thresholds (score set) in the given ranked gene list (Section 1.2.4). Thus, GSZ scores of a gene set in N permuted data are the largest values from N score sets. Assuming that these N score sets are from the same distribution, the GSZ scores from N permuted data can be described by extreme value distributions. We used type I, referred to as EVD here onwards, and the generalized EVD that combines the three types, referred to as GEVD here onwards.

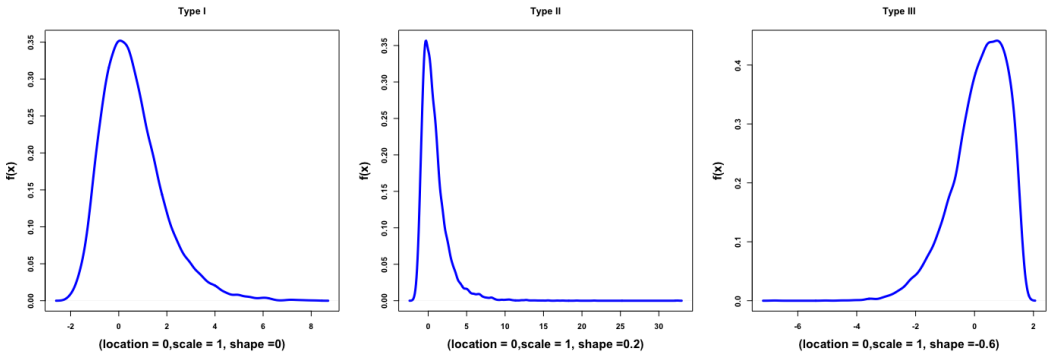


Figure 3.1: Probability density functions for type I, II and II extreme value distrubutions.

Extreme value type 1 distribution (EVD)

The EVD is defined by the following probability density function,

$$f(x) = \frac{1}{\beta} e^{-\frac{(x-\mu)}{\beta}} e^{-e^{-\frac{(x-\mu)}{\beta}}} \tag{3.1}$$

and the cumulative distribution function is given by,

$$F(x) = e^{-e^{-(x-\mu)/\beta}} \tag{3.2}$$

where, μ is the location parameter and β>0 is the scale parameter. Standard extreme value type I distribution has $\mu = 0$ and $\beta = 1$.

General extreme value distribution (GEVD)

The general extreme value distribution (GEVD) is a flexible three parameter model that combines the extreme value type I, II and III distributions. The probability density function of the GEVD is given by,

$$f(x) = \begin{cases} \frac{1}{\beta} e(-(1+az)^{-\frac{1}{a}})(1+az)^{-1-1/a} & a \neq 0 \frac{1}{\beta} e(-z - e(-z)) \\ a = 0 \end{cases} \tag{3.3}$$

.

The cumulative distribution function is,

$$F(x) = \begin{cases} e(-(1+az)^{-1/a}) & a \neq 0 \\ e(-e(-z)) & a = 0 \end{cases} \tag{3.4}$$

.

where, z = (x - μ)/β and a, β, μ are shape, scale and location parameters respectively. The scale must be positive, the shape and location can take on any real value. The sign of shape parameter tell which of the extreme value type I, II and III distributions fits the data (μ = 0: type I, μ >0: type II, μ <0: type III) (Figure 3.1).

3.1.2 P-value calculation with series expansion

EVD and GEVD based p-values were calculated from cumulative distribution function of the fitted extreme value distributions on null distribution of GSZ scores as,

$$\text{P-value}_{EVD} = 1 - CDF_{EVD}(X), \tag{3.5}$$

where CDF_{EVD} is the cumulative distribution function of EVD and X is the gene set score. The calculation of extremely small p-values is difficult as the $CDF_{EVD}(X)$ approaches 1 due to limited number precision in computing systems. This problem was solved by deriving series expansions on the logarithm of the p-value for EVD and GEVD. This series expansion in combination with asymptotic p-value estimation allows the estimation of extremely small p-values with a moderate number of permutations.

3.1.3 Reference GSA methods used in evaluation

The evaluation of the asymptotic p-value estimation method was based on comparison of the results generated with mGSZ and seven other popular GSA methods implemented by ourselves (Table 3.1). mGSZ was also compared with GSA methods with available implementations as R packages (Table 3.1).

Table 3.1: GSA methods used as reference for evaluation of asymptotic p-value estimation method and mGSZ.

Compared GSA methods	Descriptions
GSA	Maxmean statistics based method with empirical p-value estimation method implemented as R package [34].
Allez	Random-set enrichment score based method implemented as R package [88].
mGSA	Our improved version of maxmean statistics based method included in mGSZ R package.
mAllez	Our improved version of random-set enrichment score based method included in mGSZ R package.
WRS	Wilcoxon rank-sum statistics based method with empirical p-value estimation method.
SS	Sum of squares based method with empirical p-value estimation method.
SUM	Sum based method with empirical p-value estimation method.
KS	Kolmogorov-Smirnov statistics based method with empirical p-value estimation method.
wKS	Weighted Kolmogorov-Smirnov statistics based method with empirical p-value estimation method [114]
CAMERA	Correlation adjusted mean rank gene set test method implemented in R package [127]
ROAST	Rotation based gene set analysis method implemented in R package [126].

3.1.4 Datasets used in evaluation

Datasets used in evaluation of asymptotic p-value estimation methods and mGSZ were downloaded from the GSEA web site (`http://software.broadinstitute.org/gsea/datasets.jsp`). The datasets have been used as benchmark datasets in the evaluation of popular gene set analysis methods such as GSEA [114] and GSA [34].

P53 cancer data P53 is a gene in human that encodes for tumor protein p53 which is a tumor suppressor protein that prevents development of cancer cells. The P53 cancer dataset consists of genome-wide transcriptional profiles with 33 samples with a mutated p53 gene and 17 samples with the wild type p53 gene [114].

Gender data The gender dataset consists of transcriptional profiles from 15 males and 17 females lymphoblastoid cell lines [114].

Leukemia data The leukemia dataset consists of gene expression profiles of cells from 24 acute lymphoid leukemia patients and 24 acute myeloid leukemia patients [3].

3.1.5 Evaluation of asymptotic p-value estimation method

In order to evaluate the accuracy of the asymptotic p-values, we generated a "reference of truth" by calculating GSZ scores and their empirical p-values with 100000 sample permutations of the original data. Empirical p-values were calculated as the number of null GSZ scores greater than or equal to the test GSZ score divided by the total number of permutations. This is based on the assumption that empirical p-values calculated with 100000 sample permutations are closer to the true p-values. The asymptotic p-values were evaluated by calculation their Pearson correlation (cor) and mean squared error (mse) with log transformed empirical p-values. While correlation indicates the magnitude and direction of a linear relationship, mean squared error indicates the difference between asymptotic p-values (test p-values) and empirical p-values (the reference of truth). In addition to the whole list, the analysis was repeated on the signal rich region of p-values by calculating cor and mse also for the subset of p-values $<$ 0.10. This emphasizes the biologically interesting regions of the gene list. An optimality criterion was defined for the tested statistical distributions. Distributions with mse$<$0.10 and cor$>$0.97 in both the whole list and the signal rich regions were considered to be optimal distributions for asymptotic p-value estimation. The evaluation of asymptotic p-values was done with two different microarray gene expression datasets: (a) P53 cancer data (b) Gender data.

We did the same experiment with two other gene set scoring functions: GSA [34] and Allez [88]. The best statistical distribution models for p-value estimation of GSA and Allez gene set scores turned out to be EVD and NORM (gaussian), respectively. We present modified versions of GSA and Allez referred to as mGSA and mAllez respectively. mGSA includes two modifications: (1) substitution of max-mean statistics with GSZ, (2) substitution of empirical p-value estimation with asymptotic p-value estimation. mAllez includes one modification, that is, the addition of sample permutation on top of implicit gene permutation.

3.1.6 Evaluation of mGSZ: new evaluation principles for GSA methods

Literature lacked robust and unbiased evaluation principles for GSA methods. This work developed several approaches for multi-perspective evaluation of a GSA method.

Identification of relevant gene sets

Competitive gene set analysis methods aim to identify gene sets that are more associated with phenotype or "relevant" compared to a random set of genes of same size from the data. Thus, the most straightforward evaluation approach would be to score the methods based on the number of relevant gene sets identified. However, we lack knowledge of the relevant gene sets (or the truth). In this study, relevant gene sets were defined based on literature mining and transcription factor (TF) activity.

Relevant gene sets based on literature mining The top n (for example, $n = 20$ in the case of gender data and $n = 50$ in the case of p53 cancer data) most significant gene sets reported by each of the compared methods were pooled. Each of the pooled gene sets was then analyzed separately for their relevance to a condition or exposure by intensive literature mining, for example, gene sets relevant to p53 activity in P53 cancer gene expression data. Methods were compared for the number of relevant gene sets in top (for example, 20 in the case of gender data and 50 in the case of p53 cancer data) gene sets reported. The results were presented as a plot of the cumulative count of the number of relevant gene sets on the y-axis against the top n gene sets on the x-axis.

Relevant gene sets based on TF activity Gene expression data from transcription factor (TF) deletion and overexpression experiments is another useful tool for evaluating gene set analysis methods because the experimental perturbations of TFs provide the standard-of-truth for the evaluation of methods. The gene sets in such setting can be defined as targets genes of TFs. Unlike biological processes, TFs can be directly controlled and adjusted (such as overexpression or deletion) to have an expected effect on the expression levels of target genes. The evaluation is based on the ability of the compared methods to identify TFs based on changes in expression levels of their targets genes. This idea was implemented by [87] for the comparison of 14 different gene set analysis methods. The experimentally perturbed TFs were considered positive gene sets and an ideal gene set analysis method should identify the perturbed TFs. As the dataset had no biological replicates, the gene sets were ranked based on empirical p-values calculated with 100000 gene permutations. For each experiment, the area under curve (AUC) was calculated and the compared methods were ranked based on the mean and standard deviation of the AUC.

False positive test

The idea in the false positive test is to investigate if a method has a tendency to create false-positive results when gene sets are in fact not associated with the studied condition or exposure. The compared methods were run with the original gene expression data but with null gene sets generated by randomly choosing the member genes while keeping the gene set size intact. The performance of the compared methods was evaluated by examining the distribution of p-values reported from the leukemia dataset.

3.2 Advanced permutation method

Six different permutation methods (Section 3.2.1) were evaluated for the analysis of multi-group gene expression data with less than six replicates per group. All the permutation methods were implemented with GSZ. Thus, there were six different GSA methods based on the six different permutation methods. The permutation methods were evaluated based on the performance of the GSA method implementing them. As the only difference between the six methods is the type of permutation, the difference in their performance is due to the permutation method. THe evaluation of permutation methods was based on the identification of reference gene sets generated by a data split method (Section **??**) on two different datasets (Section 3.2.3). The most optimal permutation method identified after evaluation was implemented in *mGSZ*. The *mGSZ* method updated with advanced permutation method for multi-group data is referred to as *mGSZm*. *mGSZm* was evaluated by comparing its performance to seven other GSA methods (Table 3.2) in the analysis of three different multi-group gene expression datasets (Section 3.2.3).

3.2.1 Advanced permutation methods

Advanced permutation methods, in the context of this work, are meant to be used for pairwise comparisons of chosen exposures/groups. The main idea underlying the methods is to employ all the groups (including also the groups other than those being compared) for the estimation of the null distribution. The assumptions underlying the methods are that: 1) there is same inherent structure, including gene-gene correlations, across all groups, 2) the differential gene expression score represents gene regulation and co-regulation of member genes, 3) a gene set represents regulation of the associated biological process. The null hypothesis that the outcome is not related to group status must hold true with the null distribution.

In the context of GSA, if for example, one is interested in comparing a group with a particular disease with a healthy group, a test statistic under the null hypothesis should have nothing to do with the healthy vs. disease categorization. In this study, we generated a null distribution for a pair-wise comparison by permuting the compared groups as well as the other groups. In doing so, there could be a situation where highly correlated samples end up in the compared groups giving test statistics as extreme as the one obtained with original data. We call this phenomena *leakage of biological signal* which leads to false negative results (Figure 3.2). The goal of this study was to develop a permutation method for generating a null distribution of gene set scores from a pair-wise comparison in multi-group gene expression data while controlling for biological leakage.

We evaluated six different permutation methods, referred to as Perm1-6 hereafter. The methods differ from each other in the way they model the background signal of the data. Perm1 is a naive permutation method limited only to the groups being compared. Perm2-6 are candidate approaches that aim to prevent signal leakage by ensuring that the permuted groups do not contain many samples from a single original group or highly correlated groups. Identification of the most optimal permutation method for such dataset is not trivial. The degree of randomness in picking samples from groups in a dataset for permuted data plays a crucial role in the leakage of biological signal. Complete randomness can result in highly correlated samples ending up in the same group in the permuted data, thus producing false negative results. On the other hand, too constrained a method can generate false positive results. In order to find an optimal balance, perm2-6 were designed such that sampling is the least constrained in perm2 and most constrained in perm6. For example, perm2 picks samples randomly from all sample groups and can group highly correlated samples. The following notations will be used in describing the permutation methods:

n: Number of replicates in a group. For simplicity, we let $n(= 5)$ be constant for all groups.

m: Total number of groups in multi-group gene expression data ($= 6$ in our example).

y: Total number of samples in all the groups.

z: Total number of samples in the groups being analyzed.

B and D: The groups that we are interested to compare.

$B*$ and $D*$: Permutations of the analyzed groups B and D.

X: Number of permutations.

$N|Permx|$: Total number of permutations from permutation method x.

Perm1

This is the naive permutation method used in gene set analysis of two-group gene expression data. In case of multi-group data, perm1 based gene set analysis considers only the compared groups and the rest is ignored. Only the samples from the groups being compared are permuted (Figure 3.3). The number of unique permutations using perm1 is given by:

$$N|Perm1| = \frac{1}{2}\,\frac{z!}{n!\,(z-n)!} \tag{3.6}$$

The factor 1/2 is included to exclude permutations that are mirror images of one another and give identical results, such as (1,1,0,0) and (0,0,1,1). In a multi-group gene expression dataset with $m = 6$ and $n = 5$, the total number of obtainable permutations is 126. Thus, perm1 cannot generate sufficient permutations needed to construct a reliable hull distribution.

Perm2

Perm2 involves rearrangement of all the samples in the data irrespective of the pair of groups being analyzed. For example, in a data with six groups, if we are interested in analyzing groups B and D, the permuted groups $B*$ and $D*$ will include random samples from the compared as well as any of the other groups (Figure 3.4). This might allow situations where permuted groups may not contain any sample from the compared groups or may contain many or all samples from their respective original groups. Thus, there is a risk of generating an unreliable null distribution, for example, by incorporating biological signal. The total number of unique permutations by perm2 is given by,

$$N|Perm2| = \frac{1}{2}\,\frac{y!}{n!\,n!\,(y-2n)!} \tag{3.7}$$

that gives $3.8e + 9$ unique permutations from the example dataset with $m = 6$ and $n = 5$.

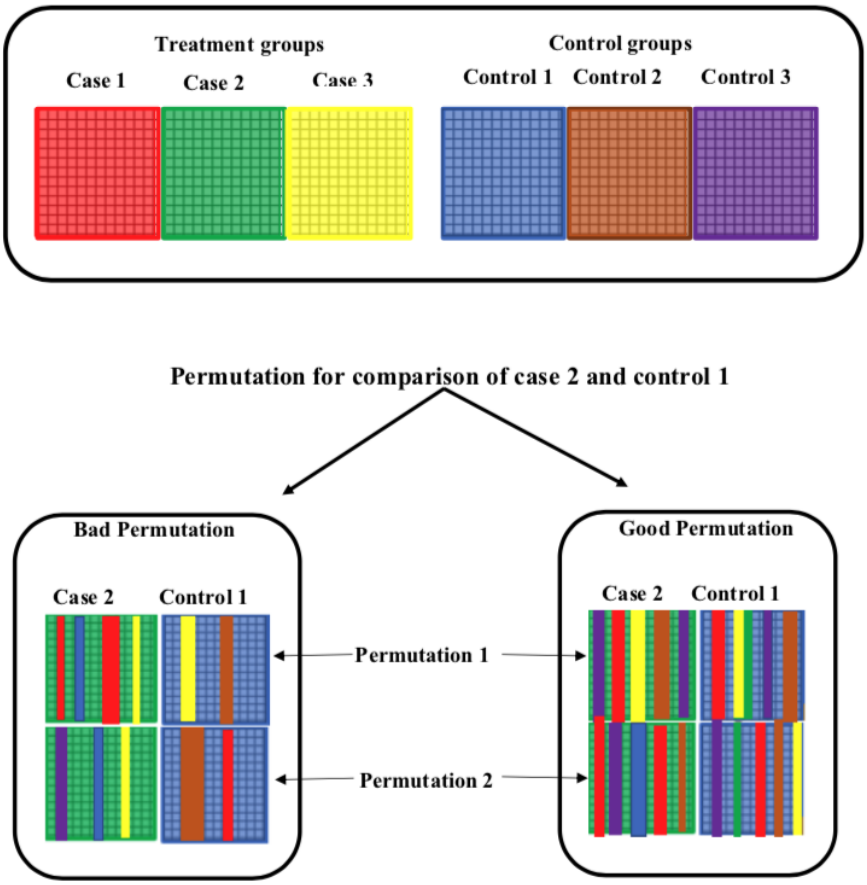


Figure 3.2: Examples of biological leakage during generation of null data. Case and control groups under the bad permutations example contain majority of replicates from similar or highly correlated sample groups, which can potentially introduce biological signal into null data. In contrast, under the good permutations example, case and control groups contain well mixed samples from all different groups in dataset.

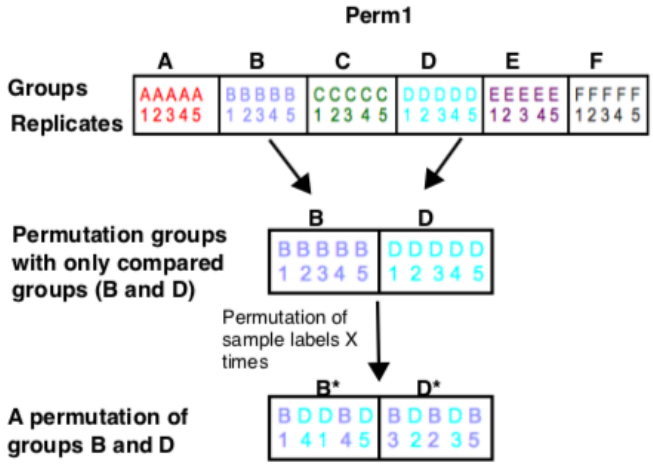


Figure 3.3: Schematic diagram representing Perm1.

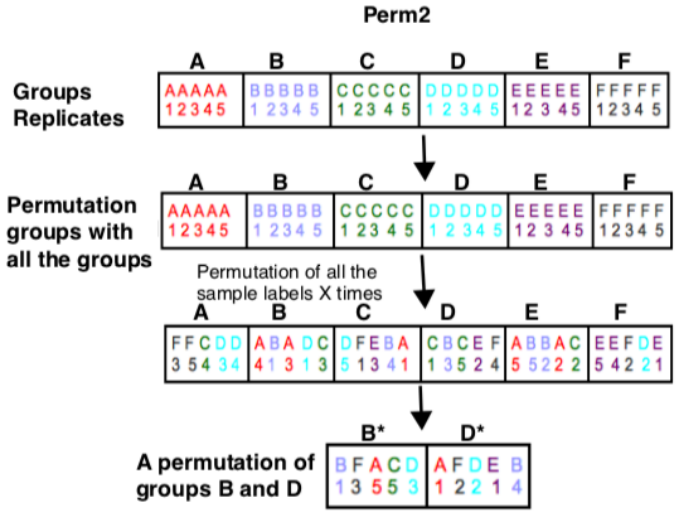


Figure 3.4: Schematic diagram representing Perm2.

Perm3

In multi-group data with $m>n$, permutations are generated in two steps: group selection followed by sample selection from those groups (Figure 1). Permutations are generated for each of the analyzed groups separately. First, n groups are randomly selected without replacement. Then sample selection is done by randomly selecting each of the n samples from n selected groups. The procedure is repeated to obtain the second permuted group, with the exception that samples found in the first permuted group are explicitly filtered out. The mean estimate is thus not strongly influenced by any individual groups. The lower bound of the number of unique permutations is given by,

$$N|Perm3| = \left(\frac{m!}{k!\,(m-n)!}\right)^2 * n^n * (n-1)^k * \frac{1}{2} \quad , m>n \tag{3.8}$$

A detailed diagrammatic illustration is presented in Appendix I.

Perm4

Perm4 is similar to perm3 except in the group selection step, where the groups to be analyzed are selected followed by sampling $n-2$ additional groups without replacement. The lower bound of the number of unique permutations is given by,

$$N|Perm4| = \frac{1}{2}\, n^n (n-1)^n \left(\frac{(m-2)!}{(n-2)!\,(m-n)!}\right)^2 \quad , m>n \tag{3.9}$$

that gives $2.6e+7$ permutations for an example case with $m = 6$ and $n = 5$.

Perm5

In this method, the permutation process for one of the analyzed groups is similar to that of Perm3. The only difference is that the permutation space for the other analyzed group is restricted to only those groups which were used for permuting the previous analyzed group (Figure 2). This prevents

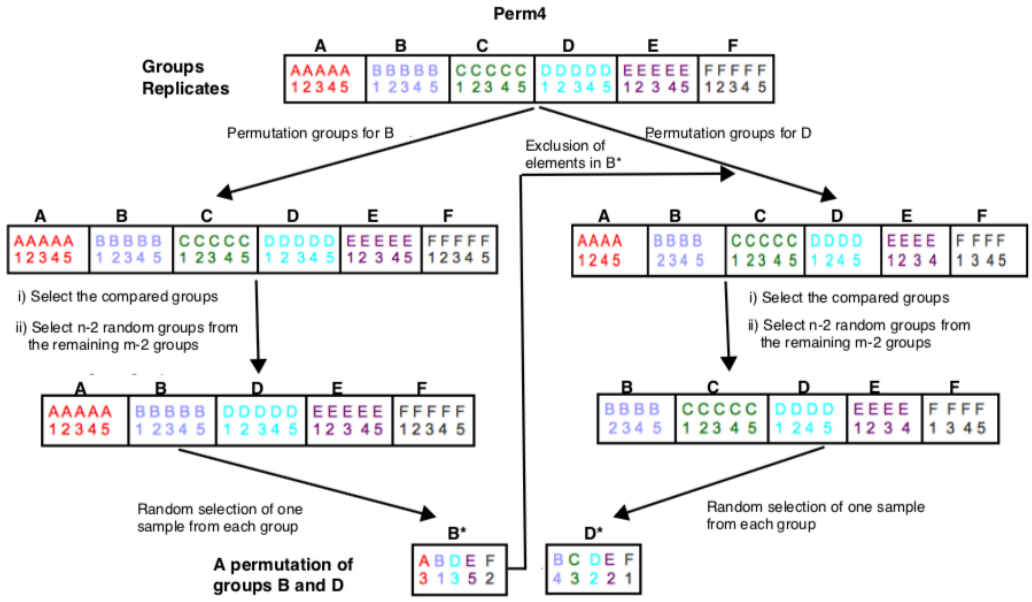


Figure 3.5: Schematic diagram representing Perm4.

leakage of biological signal to the null distribution. The total number of unique permutations for analyzed groups is given by,

$$N|Perm5| = \frac{m!}{n!\,(m-n)!} * n^n * (n-1)^n * \frac{1}{2} \quad , m>n \tag{3.10}$$

A detailed diagrammatic illustration is presented in Appendix J.

Perm6

In this method, the permutation process for one of the analyzed groups is similar to that of Perm4. The only difference is that the permutation space for the other analyzed group is restricted to only those groups which were used for permuting the previous analyzed group for the same reason as mentioned under Perm5 (Figure 3). Total number of unique permutations for analyzed groups is given by,

Table 3.2: GSA methods used as reference for evaluation of *mGSZm*

Compared GSA methods	Descriptions
romer	Rotation testing using mean ranks [103].
GAGE	Generally Applicable Gene-set Enrichment [75].
QuSAGE	Quantitative Set Analysis of Gene Expression [128].
wKS	Weighted Kolmogorov-Smirnov statistics based method with empirical p-value estimation method [114].
CAMERA	Correlation adjusted mean rank gene set test method implemented in R package [127].
Allez	Random-set enrichment score based method implemented as R package [88].
mGSA	Our improved version of maxmean statistics based method included in mGSZ R package.

$$N|Perm6 = \frac{(m-2)!}{(n-2)!\,(m-n)!} * n^n * (n-1)^n * \frac{1}{2} \quad , m{>}n \tag{3.11}$$

A detailed diagrammatic illustration is presented in Appendix K.

Permutation in exceptional cases

The derivations for perm2-6 are based on cases where the total number of sample groups is more than the number of replicates in each group. Also, the number of replicates per group was assumed to be constant for all sample groups. The derivation and explanation of a permutation method for exceptional cases is presented in Appendix L.

3.2.2 Reference GSA methods used in evaluation,

The evaluation of *mGSZm* was based on comparison of results generated with *mGSZm* and seven other popular GSA methods presented in Table 3.2.

3.2.3 Datasets used in evaluation

mGSZm was intended for GSA of multi-group data with fewer than six replicates per group. Therefore, multi-group datasets were used in the evaluation of the method. Also, as one of the evaluations was based on data splitting where we use the method to analyze smaller (test) and larger

(reference) subsets of the data and evaluate the method based on its performance on test subset as compared to reference subset (Section **??**), two of the used datasets have a minimum of 15 biological replicates.

Human primary cell data

Human primary cell gene expression data was downloaded from the Gene Expression Omnibus data repository (GEO accession: GSE49910). A total of 124 arrays of Affymetrix Human Genome U133 Plus 2.0 expression arrays were downloaded. The dataset consists of eight sample groups of different cell types - embryonic stem cells, tissue stem cells, epithelial cells, fibroblasts, endothelial cells, osteoblasts, keratinocytes and smooth muscle cells. The evaluation of the methods was based on analysis of the sample groups that have at least 15 biological replicates, are well clustered and have no outliers (endothelial cells and keratinocytes).

Breast cancer data

Breast cancer data was downloaded from the Gene Expression Omnibus data repository (GEO accession GSE3165). The data includes 94 arrays of platform GPL887 (Agilent Human 1A Microarray V2) with six sample groups corresponding to six molecular subtypes of breast cancer. The subtypes are basal-like, luminal A, luminal B, Her2, normal-like and claudin-low. The evaluation of the methods was based on comparison of the sample groups that have at least 15 biological replicates, are well clustered and have no outliers (Basal-like and Her2).

Mouse tissue gene expression data

Mouse tissue specific gene expression data was downloaded from the Gene Expression Omnibus data repository (GEO accession GSE9954). The data is based on Affymetrix Mouse Genome 430 2.0 Array and consists of expression profiles from six mouse tissues (kidney, liver, lung, heart, muscle and adipose). The dataset consists of four replicates for kidney and three replicates for each of the rest of the tissues.

Gene sets

Curated gene sets from the Molecular Signatures Database were used [114].

3.2.4 Evaluation of permutation methods

The evaluation of permutation methods was done with a novel approach based on data splitting, where the smaller subset is used to test methods and the bigger subset (the rest of the data) is used to generate reference gene sets.

Identification of reference gene sets based on data splitting

The data splitting method partitions a big gene expression dataset into test and reference subsets (Figure 3.6) and thus requires a dataset with a large number of replicates per group. After splitting, the test and reference partitions comprise 25 % and 75 % of the data (arrays), respectively. Reference gene sets were generated by taking the union of the top most significant gene sets (for example, 5) reported by each of the compared methods for the reference data. Possible bias due to the selection of a number of top gene sets was minimized by repeating the procedure for multiple numbers of top most significant gene sets, for example, 3,5 and 7. The evaluation of the compared gene set analysis methods was based on the cumulative count of reference gene sets in the top, for example, 20 gene sets returned by each of the methods. The cumulative counts at each rank were the average of counts obtained across different runs with a different number of top gene sets of reference data. The procedure was repeated multiple times (for example 100 depending on the test) for different data splits and the cumulative counts were averaged over all the runs. The results were presented as a plot of the averaged cumulative counts.

Permutation methods were implemented with GSZ and their evaluation was based on the ability of the methods to identify reference gene sets generated with data splitting method described above using human primary cell data and breast cancer data (Section 3.2.3).

3.2.5 Evaluation of mGSZm

mGSZm was evaluated with three different methods:

Identification of reference gene sets based on data splitting

Similar to the evaluation of permutation methods, *mGSZm* was also evaluated based on its ability to identify reference gene sets generated with data

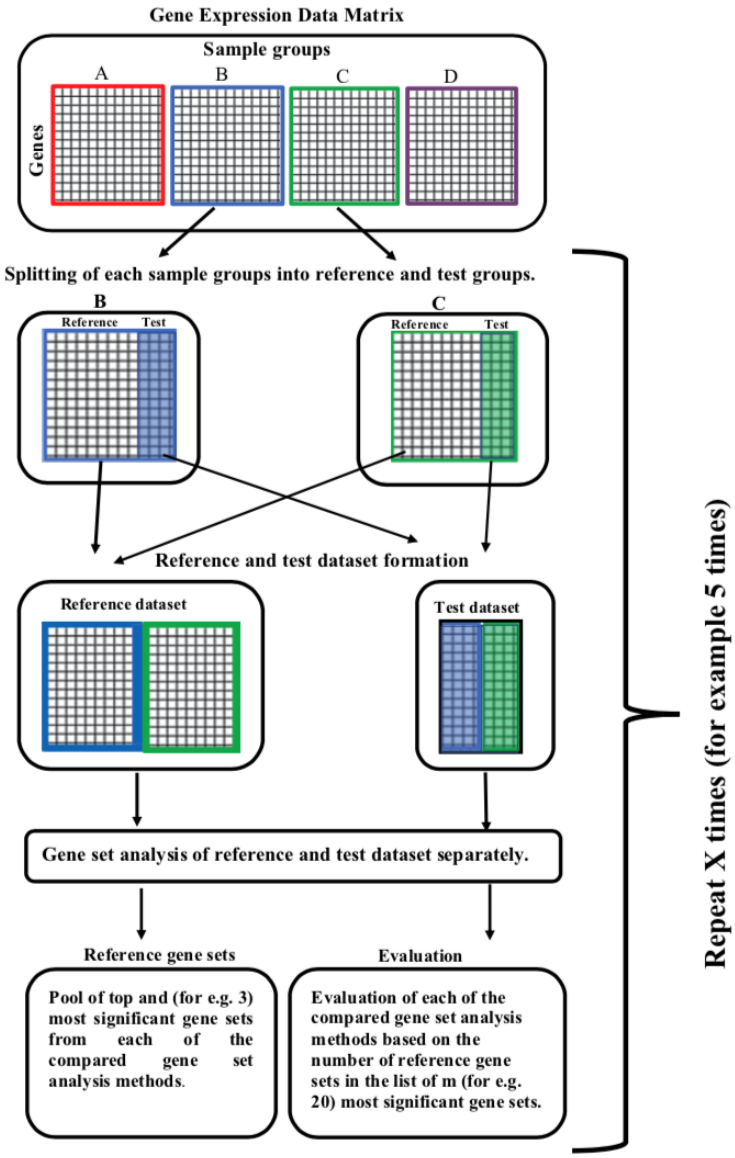


Figure 3.6: Evaluation based on data splitting. Workflow of the evaluation based on splitting of data.

splitting method using human primary cell data and breast cancer data (Section 3.2.3)

Identification of tissue specific gene sets

Genes that are functional and thus expressed in specific tissues only are called tissue specific genes. Tissue specific genome-wide gene expression data can be a useful tool for evaluating gene set analysis methods, because a considerable number of studies on tissue specific gene expression exist [80], [25]. Tissue specific gene sets were generated using tissue specific genes identified and verified by previous studies [110]. As gene sets available from various sources might contain a certain amount of noise (irrelevant genes), different proportions of noise were also introduced to the tissue specific gene sets. This was done by randomly selecting x% (where x(0,10,20,30,40,50,60,70,80,90)) of genes in a tissue specific gene set and replacing them with randomly selected genes from the remaining data. This procedure gives 10 gene sets for each tissue type with noise level varying from 0% to 90%. These gene sets can then be mixed with randomized gene sets containing other genes in the analyzed dataset. *mGSZm* was evaluated based on its ability to rank tissue specific gene sets higher than the random gene sets as compared to reference GSA methods. Pairwise comparisons between all possible tissues were done and results were presented as the average cumulative count of tissue specific gene sets in top 50 gene sets.

False positive test

mGSZm was evaluated for its ability to control the false positive rate (type 1 error) by investigating the p-value distribution of gene set scores from null gene expression data. Null gene expression data was generated by randomizing the sample labels of the analyzed gene expression data. P-values estimated by mGSZm or similar methods based on null gene expression data with no true differential gene expression should follow a uniform distribution.

3.3 Gene Set Analysis of genome-wide methylation data

Gene set analysis has become standard practice among various omics methods [85]. However, for datasets other than gene expression, such as DNA methylation data, over-representation analysis is more popular [53, 129, 92, 23]. The approach involves the generation of a list of the most interesting

CpG sites based on some threshold such as p-values for differential methylation scores. The list of CpG sites is reduced to a list of genes with several ad hoc methods such as including genes with one or more CpG site(s) annotated to the gene significantly associated with respective studied phenotypic trait [104]. One of the major limitations of these methods is that the results vary with varying thresholds used to select the list of the most interesting CpG sites. Another major limitation of over-representation analysis of methylation data is that genes with a larger number of CpG sites within their genomic region have a higher probability of having at least of CpG with significant difference in methylation based on a threshold. Such genes are more likely to end up in a list of differentially methylated genes and consequently lead to biased pathway analysis results. This phenomenon has been shown by [39].

The problem of ambiguity in results due to ambiguous thresholds used to select the CpG list is addressed by hunctional class scoring methods based on signal summary and ranked list, as whole genome-wide DNA methylation at gene level data is analyzed (Section 1.2.3). CpG level genome-wide DNA methylation data can be summarized to gene-level genome-wide DNA methylation by several approaches, such as taking the average of the methylation levels of all CpGs mapping to the genomic region of a gene or using the methylation level of a CpG with maximum difference in methylation level between compared groups as a proxy for gene level methylation. My third contribution is the application of such methods for gene set analysis of methylation data to investigate epigenetic signatures of smoking at biological pathway level (Publication III). Gene set analysis of DNA methylation data shifts the analysis from the level of individual CpG sites to gene sets level. The outcome is the understanding of potential biological significance of the methylation changes caused by smoking. This approach enhances the statistical power of test of association between smoking and methylation by pooling methylation levels of a set of genes linked to the same biological pathways. The GSZ-based gene set analysis method published in Publication I was applied for the analysis. We analyzed gene sets from the Molecular Signature Database (MSigDB), a collection of gene sets from various annotation libraries such as Gene Ontology [5], KEGG pathways [56] and reactome [28].

CpGs were mapped to biological pathways via gene annotations. CpGs mapping to any genomic region of the corresponding genes (TSS200, 0-200 bases upstream of the transcriptional start site; TSS1500, 200-1500 bases

upstream of the TSS; 5'UTR, within the 5' untranslated region, between the TSS and the ATG start site; Body, between the ATG and stop codon irrespective of the presence of introns, exons, TSS, or promoters; 3'UTR, between the stop codon and poly A signal) were considered as candidate proxy for gene level methylation level. As CpG sites from different regions of a gene were considered as proxy for the gene level methylation, speculation on whether the altered methylation activates or deactivates gene expression is inconclusive and thus outside the scope of this work. However, it is important to note that the function of DNA methylation varies with different genomic contexts [54].

3.3.1 Study population

The Cardiovascular Risk in Young Finns study (YFS) is a prospective multi-centre study initiated in 1980 (number of subjects=3596, baseline age 3-18 years). The participants have been followed up over 40 years to investigate childhood risk factors for cardiometabolic outcomes in adulthood [98]. The follow-up includes comprehensive data collection using questionnaires, physical measurements, dietary interviews and blood tests from childhood to young adulthood. This study was based on a subset of 192 participants for whom DNA methylation measurements from whole-blood samples were available from the 2011 follow-up. The smoking history of the participants was self-reported and belonged to six categories based on smoking frequency (1. active smoker or at least once a day, 2. once a week or more often, however not daily, 3. less often than once a week, 4. attempts to quit, 5. has quit, 6. has never smoked). To obtain maximum effect size, the study was focused on sub-sample of 125 participants, 40-49 years of age who were either active smoker (n=21) or have never smoked (n=104). The study has been approved by the ethical committee of the Hospital District of Southwest Finland and the Regional Ethics Committee of the Expert Responsibility area of Tampere University Hospital. All subjects have given informed written consent.

3.3.2 DNA methylation assessment

DNA was extracted from 192 EDTA-blood samples using WizardÂ®Genomic DNA Purification Kit (Promega Corporation, Madison, WI, USA) according to the manufacturer's instructions. Genome-wide quantification of DNA methylation levels was done using Illumina Infinium HumanMethylation450

BeadChips [15], [14], [13] in the Core Facility at the Institute of Molecular Medicine Finland (FIMM), University of Helsinki following manufacturer's protocols. The HumanMethylation450 BeadChip measures DNA methylation at more than 485,000 CpG cites across the genome. The arrays were imaged with a high-precision scanner (iScan system, Illumina Inc.), and the signal intensities were extracted using a software package (GenomeStudio Software, Illumina Inc.).

3.3.3 Data pre-processing

Data was obtained and processed from raw methylation image files using the minfi package in R/Bioconductor [4]. Samples with the sum of detection P-values across all the probes above 0.05 were excluded from the analysis. Samples in which the log2 median of methylated and unmethylated intensities did not cluster within the default cutoff (10.5) in the getQC function in minfi were filtered out. Further, samples for which the real sex did not match the predicted sex with the getSex function in minfi were excluded. Background subtraction and dye-bias normalization were performed via the noob method [120] implemented in minfi. Further, stratified quantile normalization was performed using the preprocessQuantile function in minfi. Probes with detection P-value above 0.01 in 99% of the samples were filtered out. CpG loci on sex chromosomes were excluded from the analysis to avoid gender based methylation bias. Also, cross-reactive probes and probes with single nucleotide polymorphisms (SNPs) were excluded from the analysis. After quality control, the total number of autosomal CpGs was 429,773. Similarly, the total number of samples was reduced to 186 which included 21 active smokers, 104 non-smokers, 58 former smokers and 3 with no smoking information. As the focus of this study was to investigate pathway level methylation changes induced by smoking, we excluded former smokers from the analysis for maximal effect size. Batch correction based on control probes removes technical artifacts in the data. This was done by adding first two principal components of control probes as covariates in the linear model during CpG-wise differential methylation analysis between active smokers and non-smokers.

3.3.4 Gene set analysis

Differential methylation analysis

All statistical analyses were performed using R (v.3.3.2) [97] based on M-values, the log2 ratio of the intensities of methylated probe versus unmethylated probe. Differentially methylated CpG loci with respect to smoking

status were identified using multivariate linear regression implemented in the CpGassoc R package [8]. All analyses were adjusted for age, sex, bmi, technical covariates (chip and array), white blood cell type proportions and the first two principal components of array control probes. The fraction of white blood cells (CD8T, CD4T, NK cells, B cells, monocytes, and granulocytes) was estimated through the reference-based Houseman method [49] using the estimateCellCounts function in the minfi package. Adjustment for cell composition was done by including the six estimated cell type fractions as covariates in the multivariate linear regression. Differential methylation scores (t-statistic) for genes were identified by utilizing as a proxy the CpG site with maximum absolute t-score from any genomic regions in relation to the genes. Since the mechanism how methylation influences gene expression is not completely understood, we limit this study only to the identification of smoking related methylation patterns within the genomic region of a gene, be it coding or promoter region. The annotation database provided by Illumina was utilized to connect the CpG sites to the genomic identifiers [46].

mGSZ analysis

Gene set analysis was performed to understand smoking induced alteration in DNA methylation at biological process or pathway level. The analysis was performed using the gene set analysis method implemented in *mGSZ* package in R [82]. Similar to gene set analysis of gene expression data, gene-gene correlation needs to be considered also in DNA methylation data analysis as CpGs are correlated just like genes are. The correlation among CpGs have been shown to increase as the proximity between CpGs increase [111]. *mGSZ* estimates a p-value for gene set scores based on the permutation of sample labels instead of genes. This approach keeps the gene-gene correlation structure intact and thus prevents false positive results (Section 1.2.5). GSA methods based on gene permutation assume that genes are independent, leading to spurious results [41]. Gene sets were downloaded from the Molecular Signatures Database (MSigDB) [114], [69]. The database contains a collection of 22569 gene sets as of September 21, 2019, categorized into eight major categories based on sources and method of generation: hallmark, curated, positional, motif, computational, gene ontology, oncogenic and immunogenic signatures.

Chapter 4

Results and discussion

4.1 Asymptotic p-value estimation

This study introduces an asymptotic method of p-value estimation for competitive gene set analysis in Publication I [82]. The asymptotic p-value is estimated by using an approximation of the true distribution. In the case of GSA, the distribution of gene set scores calculated from set of all possible permutations represents the true null distribution. The P-value is then estimated from the cumulative distribution function of the fitted distribution. This approach requires fewer permutations to estimate p-values with better resolution than 0.001 as returned by popular permutation based GSA methods. Thus, the approach significantly speeds up the gene set analysis process. Knijnenburg et al., [59] proposed a similar method for p-value calculation. It, however, requires more permutations than our proposed method.

Publication I investigates whether or not GSZ scores can be modelled by any known statistical distributions. We chose the extreme value distribution (EVD), general extreme value distribution (GEVD), gamma and normal distributions as candidates. The gamma and normal distributions were chosen as controls. GSZ scores for each gene set were calculated 500 times with 500 sample (or array) permutations of the original data. Sample permutations were done to nullify the biological effect in the data. The distributions were fitted to the empirical null distribution of the gene set scores. Distributional parameters were estimated from the fitted models and P-values were calculated from the cumulative distribution function. We refer to the p-values calculated from the fitted distributions as asymptotic p-values.

This section presents results from the evaluation of asymptotic p-values with two datasets: 1) P53 cancer data, and 2) gender data. Asymptotic p-values calculated with 500 sample permutations were evaluated against empirical p-values calculated from 100000 sample permutations (the reference of truth) (Section 3.1.5). Empirical p-values calculated with 500 sample permutations were used as a negative control. Results are presented as negative log10 p-values and thus the higher the value, the lower the pvalue. This transformation was done in order to make visualization of the plots easier.

4.1.1 Evaluation of asymptotic p-values

Results from p53 cancer data

Empirical p-values calculated with 500 permutations failed to accurately determine p-values smaller than 1/500 (green dots in Figure: 4.1) because the minimal obtainable empirical p-value depends on the number of permutations. P-values obtained from a Gaussian model fitted to GSZ scores failed suggesting that the model is not appropriate for GSZ scores which are extreme values (black dots in Figure: 4.1). P-values obtained from the extreme value (EVD) and general extreme value (GEVD) distributions fitted to mGSZ scores were the best estimates (red and blue dots in Figure 4.1), with the best correlation and mse scores (Table 4.1). This is expected because, based on the calculation, GSZ is the highest absolute value in a profile generated from subsets of ranked gene list by applying subsequent threshold at each gene position (Section 1.2.4). EVD was chosen as the most optimal model for the null distribution of GSZ scores, because the mean squared error when compared with the "reference of truth" was the smallest for EVD (Tables 4.1). Thus, the results suggest that EVD models the null distribution of GSZ scores from permuted data accurately, allowing the end users to make accurate p-value estimation with data having as few as three biological replicates.

Similarly, in the case of mGSA, EVD turned out to be the best model (Appendices A and B). As expected in the case of mAllez, which is sum based gene set score, NORM is the best model (Appendices A and C). Notice that empirical p-values based on 500 permutations fail with mGSA and mAllez.

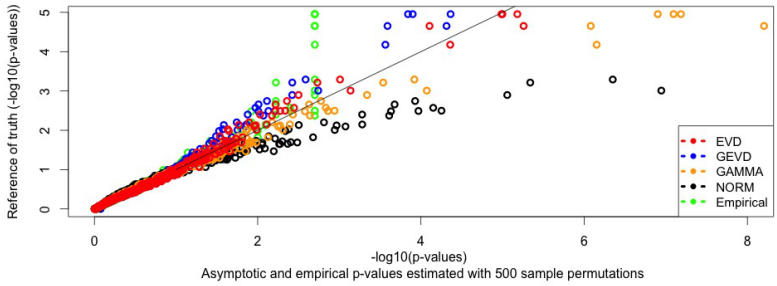


Figure 4.1: Scatter plot of asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations against the reference p-value (Y-axis). The reference p-value corresponds to empirical p-values estimated from 100000 sample permutations. Green dots are empirical p-values calculated with 500 sample permutations and represent negative control. Data - P53 cancer data. Abbreviations: EVD, Extreme Value type I Distribution; GEVD, General Extreme Value Distribution; GAMMA, Gamma distribution ; NORM, Normal distribution; Empirical, Empirical distribution.

Table 4.1: Correlation (*cor*) and mean squared error (*mse*) of asymptotic p-values estimated from statistical distribution models fitted on the null distribution of GSZ scores generated with 500 sample permutations against reference p-values generated with 100000 sample permutations with p53 data. The analysis was performed with whole gene sets list as well as subset of biologically interesting regions (p -values < 0.10). Scores not meeting the criteria (Section 3.1.5) for optimal model are highlighted. Abbreviations: EVD, Extreme Value type I Distribution; GEVD, General Extreme Value Distribution; GAMMA, Gamma distribution ; NORM, Normal distribution; Empirical, Empirical distribution; mse, mean squared error; cor, Pearson correlation.

Method	**Models**	***mse***	***mse(subset)***	***cor***	***cor(subset)***
mGSZ	EVD	0.005	0.03	0.99	0.99
	GEVD	0.02	0.09	0.99	0.99
	GAMMA	0.07	**0.36**	0.98	0.98
	NORM	**1.68**	**10.65**	**0.89**	**0.95**

Results from gender data

The lowest GSZ score p-value obtained from gender data is ˜0.001. Whereas, the one with p53 cancer data is ˜0.00001. This difference between the datasets is interesting as it clearly illustrates when the empirical approach for p-value estimation fails. For example, in contrast to the results from p53 data, empirical p-values calculated with 500 permutations are equally good as asymptotic p-values calculated with EVD and GEVD (Figure: 4.2) (Table 4.2). This is because the lowest p-value is achievable already with 500 sample permutations. Similar results were observed with mGSA and mAllez (Appendices D, E and F).

4.1.2 Evaluation of mGSZ

The performance of mGSZ was evaluated by comparing results generated by mGSZ with that generated by state-of-the-art methods that are available as R packages.

Comparison of mGSZ with program packages

The performance of mGSZ was compared to that of GSA [34]), Allez ([88], CAMERA [127] and ROAST [126]. We also included mGSA and mAllez in the comparison to show the improved performance as compared to their

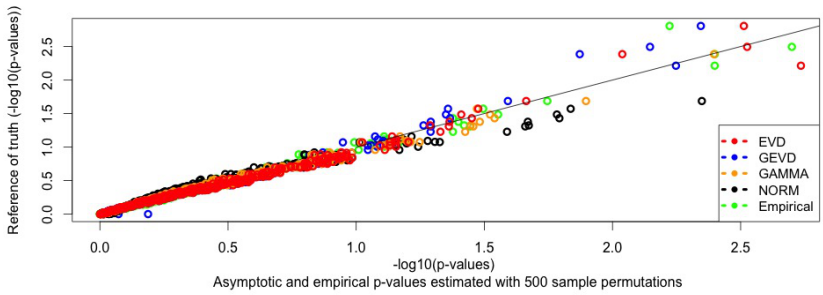


Figure 4.2: Scatter plot of asymptotic p-values (EVD, GEVD, GAMMA and NORM) (X-axis) estimated from 500 sample permutations against the reference p-value (Y-axis). The reference p-value corresponds to empirical p-values estimated from 100000 sample permutations. Green dots are empirical p-values calculated with 500 sample permutations and represent negative control. Data - Gender data. Abbreviations: EVD, Extreme Value type I Distribution; GEVD, General Extreme Value Distribution; GAMMA, Gamma distribution ; NORM, Normal distribution; Empirical, Empirical distribution.

Table 4.2: Correlation (*cor*) and mean squared error (*mse*) of asymptotic p-values estimated from statistical distribution models fitted on the null distribution of GSZ scores generated with 500 sample permutations against reference p-values generated with 100000 sample permutations with gender data. The analysis was performed with whole gene sets list as well as subset of biologically interesting regions (p -values < 0.10). Scores not meeting the criteria for optimal model are highlighted. Abbreviations: EVD, Extreme Value type I Distribution; GEVD, General Extreme Value Distribution; GAMMA, Gamma distribution ; NORM, Normal distribution; Empirical, Empirical distribution; mse, mean squared error; cor, Pearson correlation.

Method	**Models**	***mse***	***mse(subset)***	***cor***	***cor(subset)***
mGSZ	EVD	0.002	0.03	0.99	0.99
	GEVD	0.002	0.03	0.99	0.99
	GAMMA	0.003	0.05	0.99	0.98
	NORM	0.03	**0.73**	**0.96**	0.99

original versions, GSA (Efron and Tibshirani, 2006) and Allez (Newton et al. , 2007). The comparison was based on three tests: (i) Detection of relevant gene sets, (ii) False-positive signal test and (iii) P-value test where we compare the log p-values reported by each of the compared methods.

We summarize the results in Table 4.3. Two datasets (p53 and gender) were used for all evaluation tests except for false positive signal analysis. For false positive signal analysis, the leukemia dataset was used. As 79.9 % of the individual genes are differentially expressed in the leukemia data [30], it is an ideal data to test the tendency of methods to generate false positive results. Competitive gene set analysis methods compare genes in a gene set with genes not in gene sets. So, a large proportion of differentially expressed genes should not result in a large number of differentially expressed genes sets. CAMERA showed the best results in false positive signal analysis. However, in the remaining two evaluation tests, mGSZ outperformed the other methods.

False positive signal test The false positive signal test shows that mGSZ and most other methods show quite similar behavior (Figure 4.3). While CAMERA is the most conservative, ROAST shows a strong noise signal (Figure 4.3). The result points out a major difference between competitive and self-contained gene set analysis methods. Self-contained gene set analysis methods calculate the gene set scores without considering

Table 4.3: Summary of results from various evaluation tests performed for comparison of mGSZ, mGSA, mAllez with the program packages. Numbers indicate approximate rank of the methods based on the test results. Asterisk sign indicates that a test is not applicable to a particular method.

Evaluations	Data	mGSZ	mGSA	GSA	mAllez	Allez	CAMERA	ROAST
Detection of relevant gene sets	p53	1	2	3	6	7	4	5
	Gender	1	4	6	5	7	2	3
P-value resolution	p53	1	2	6	3	*	4	5
	Gender	1	3	4	2	*	5	6
False positive signal analysis	Leukemia	2	2	2	2	*	1	6

genes other than member genes. Thus, in a signal rich dataset like the leukemia dataset, most of the null gene sets are reported as significant by self-contained gene set analysis method like ROAST.

Detection of relevant gene sets The evaluation was done with p53, gender and TF datasets. In the case of the p53 dataset, the top fifty gene sets reported by each of the compared methods from P53 were pooled. Each of the pooled gene sets was then analyzed separately for their relevance to p53 activity. A total of forty gene sets highly relevant to p53 activity were selected as relevant gene sets for p53 data (Appendix G). In the case of the gender dataset, the top twenty gene sets reported by each of the compared methods were pooled. Each of the pooled gene sets were then analyzed separately for their relevance to gender. A total of ten gene sets highly relevant to gender were selected as relevant gene sets for gender data (Appendix H). Methods were compared for the number of relevant gene sets in the top n (50 in P53 data and 20 in gender data) gene sets reported. The results are presented as a cumulative count plot. The plot represents the number of relevant gene sets on y-axis and top n gene sets on x-axis. Based on results from p53 and gender data with mGSZ, mGSA, mAllez and the program packages, mGSZ is clearly the best method (Figure 4.4a and 4.4b). Moreover, mGSA and mAllez show improved performance as compared to GSA and Allez (Figure 4.4a and 4.4b).

P-value test mGSZ reports the best p-values (based on resolution) with the p53 dataset for the top 50 gene sets as compared to the other methods (Figure 4.5a). However, in case of the gender dataset, mGSZ reports the best p-values for the upper region of the gene list and then slightly lags behind mGSA, mAllez and ROAST in the lower region of the gene list (Figure 4.5b).

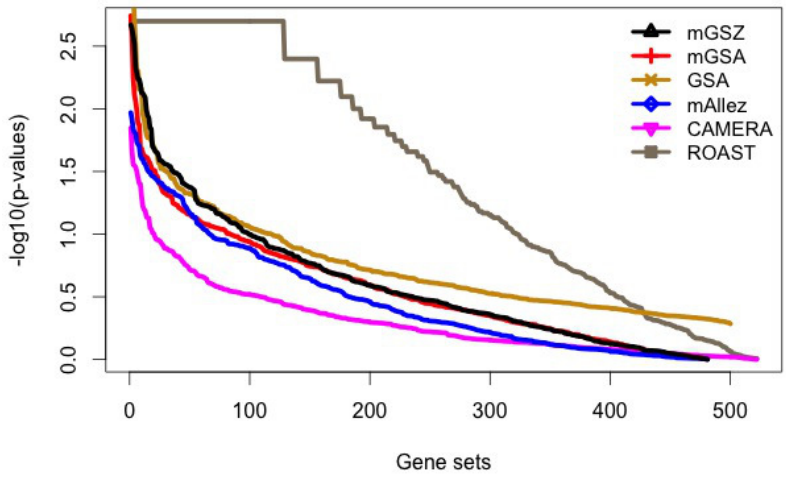
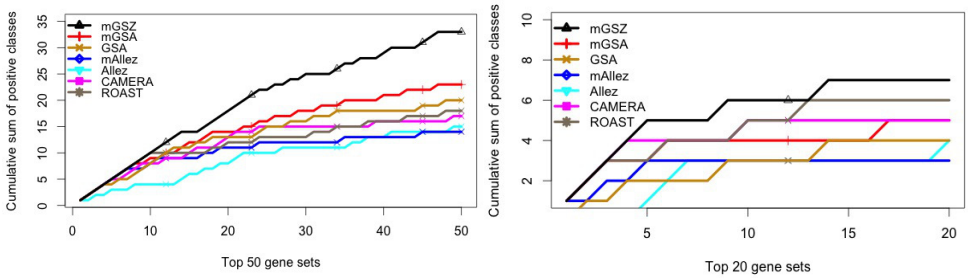


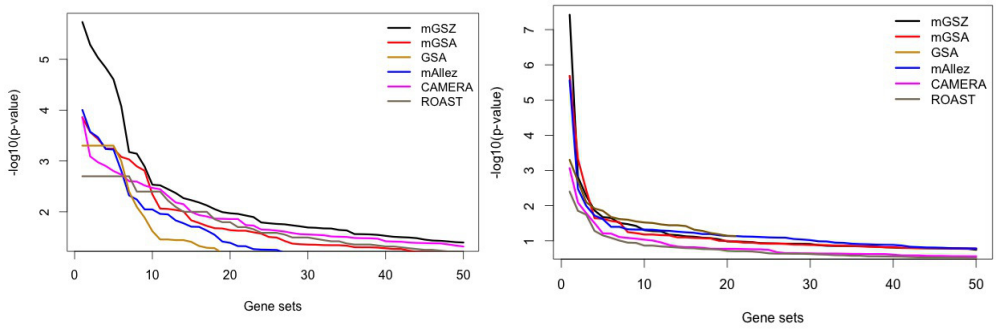
Figure 4.3: Comparison of mGSZ with the program packages with randomized gene sets.



(a) P53 data

(b) Gender data

Figure 4.4: Relevant gene sets identified by the compared methods. Figures represent cumulative count of biologically relevant gene sets (Apendices G and H) over the ranked list of top 50 gene sets in case of the p53 data and top 20 gene sets in case of the gender data reported by each of the compared methods. mGSZ (black) shows the best performance.



(a) P53 data

(b) Gender data

Figure 4.5: Log p-values for the top 50 gene sets estimated by each of the compared methods in p53 and gender datasets.

4.2 Advanced Permutation method

The problem of accurate p-value estimation with a smaller subset of permutations was addressed by introducing an asymptotic p-value estimation method in Publication I [82]. However, at least six replicates per group is recommended for reliable results. With smaller data, the number of obtainable permutations is not sufficient to generate an accurate null distribution. Multi-group (meaning data with > 2 groups in this thesis) gene expression data with fewer than six replicates are common due to various resource constraints. Naive application of sample label permutation for p-value estimation is potentially unreliable. We developed and evaluated advanced permutation methods for gene set analysis of such datasets (Publication II).

4.2.1 Evaluation of permutation methods

Perm3-6 showed similar performance in evaluations and therefore, for clarity, we only present results for Perm 4 in detail in this chapter. Perm1, the naive permutation method was the worst performing method with both datasets (Figure 4.6). Perm4 reported ~ 8 (average over 20 different data splits) more relevant gene sets at rank positions 27 to 33 in breast cancer data and 46 to 50 in primary cell data. Perm2 showed similar results, however, we prefer Perm4 to Perm2 because Perm2 is the least constrained permutation method and it cannot prevent biological signal leakage into

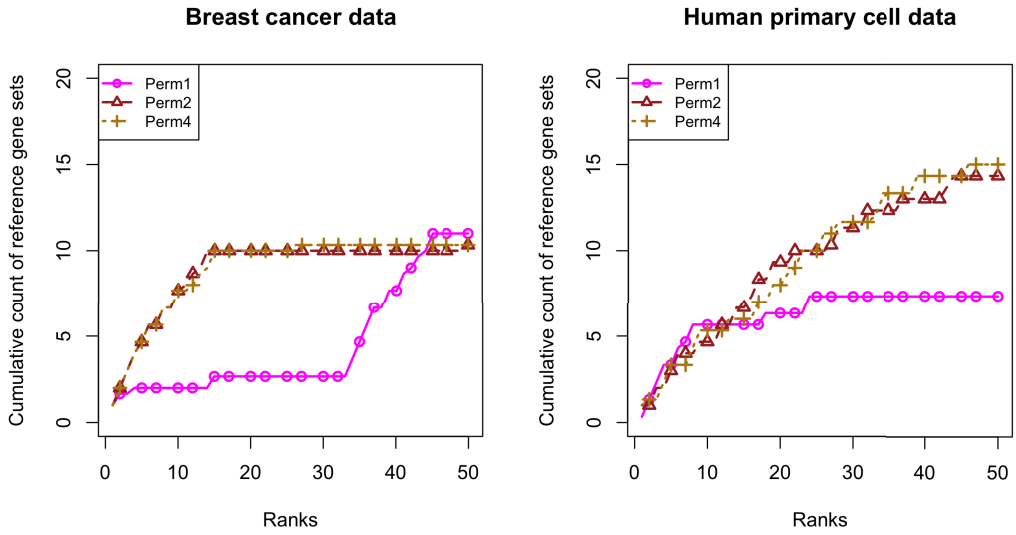


Figure 4.6: Plot showing cumulative count of reference gene sets (y-axis) across ranked list of top 50 gene sets (x-axis) for evaluation of permutation methods. Results from three gene set analysis methods based on three different permutation methods with two different datasets; 1) Breast cancer data, 2) Human primary cell data are presented.

the null distribution.

4.2.2 Evaluation of mGSZm

Perm4 was selected as the most appropriate permutation method based on evaluation presented in Section 4.2.1. Thus, *mGSZm* is based on implementation of perm4 in *mGSZ*. We evaluated the new method by comparing its performance with seven popular gene set analysis methods (Table 3.2) using three different evaluation methods (Section 3.2.5).

Identification of relevant gene sets

mGSZm was compared with seven other methods based on the identification of reference gene sets generated with the data splitting method explained in Section **??**. Results from both breast cancer and human primary cell datasets rank mGSZm as the best method (Figure 4.7). The performance of CAMERA is close to that of mGSZm in Breast cancer and QuSAGE has similar performance to that of mGSZm in human primary

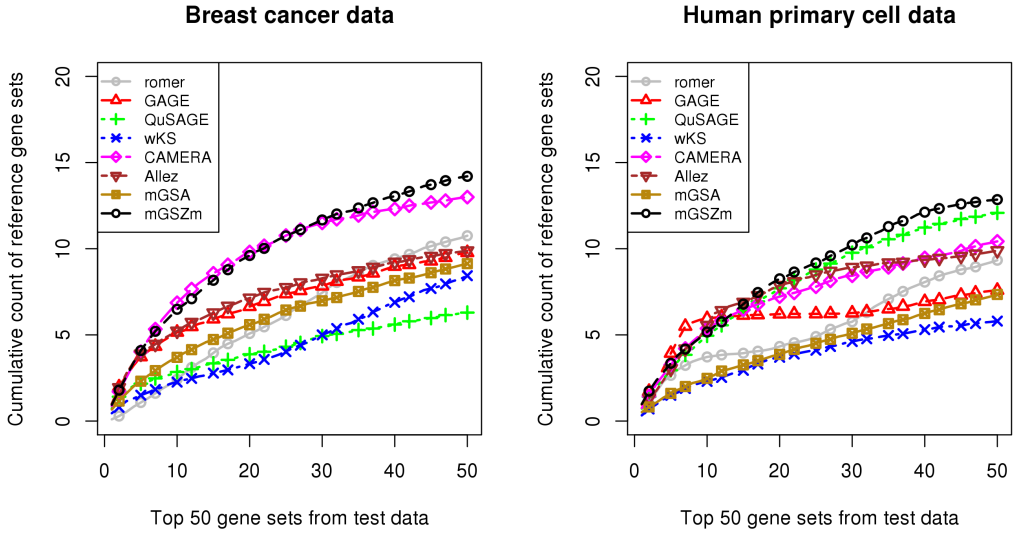


Figure 4.7: Comparison of mGSZm with seven other popular gene set analysis methods based on the cumulative count of reference gene sets (y-axis) identified across the ranked list of top 50 gene sets (x-axis). Results from two different datasets are presented; 1) Breast cancer data, 2) Human primary cell data.

cell data. However, while performance of mGSZm is consistent with both datasets, CAMERA and QuSAGE performs better in one but fails in the other dataset (Figure 4.7). GAGE reported about one gene set more than mGSZm at rank positions 4 to 15, however the performance of the methods dropped clearly at the rest of the rank positions in primary cell data.

Identification of tissue specific gene sets

mGSZm was also evaluated with seven other gene set analysis methods based on its ability to identify tissue specific gene sets as explained in Section 3.2.5. mGSZm, Allez and QuSAGE performed similarly well in this evaluation test (Figure 4.8). Inconsistencies in performance of the other compared methods are also clear in this evaluation. For example, CAMERA, which was a close competitor to mGSZm in other evaluations, is one of the worst performers in this evaluation. Further, Allez and QuSAGE, which are close competitors to mGSZm in this evaluation, performed weakly in other evaluations (Figure 4.7).

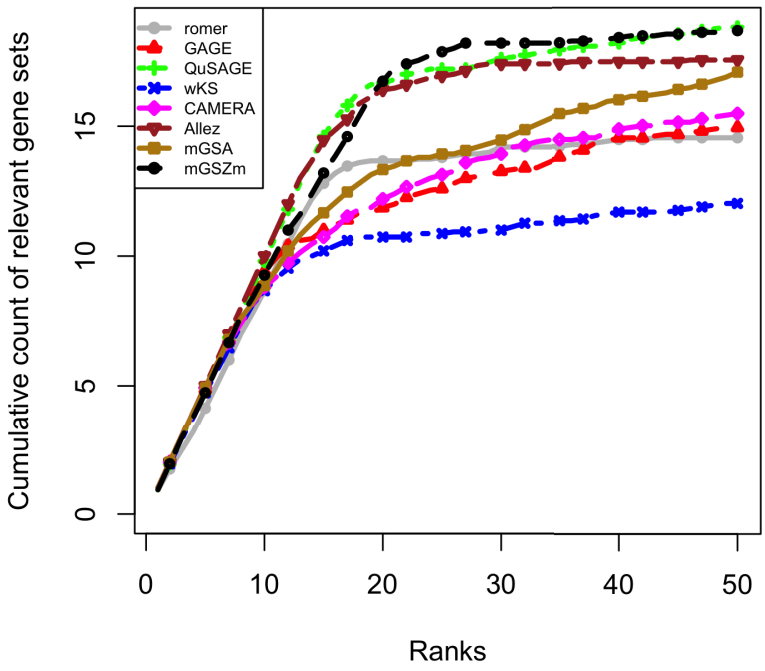


Figure 4.8: Cumulative count of tissue specific gene sets (y-axis) across ranked list of top 50 gene sets returned by eight compared gene set analysis methods. Figure represents average cumulative count of tissue specific gene sets over the ranked list of top 50 gene sets reported by each of the compared methods in 15 pairwise comparisons of six different tissue samples.

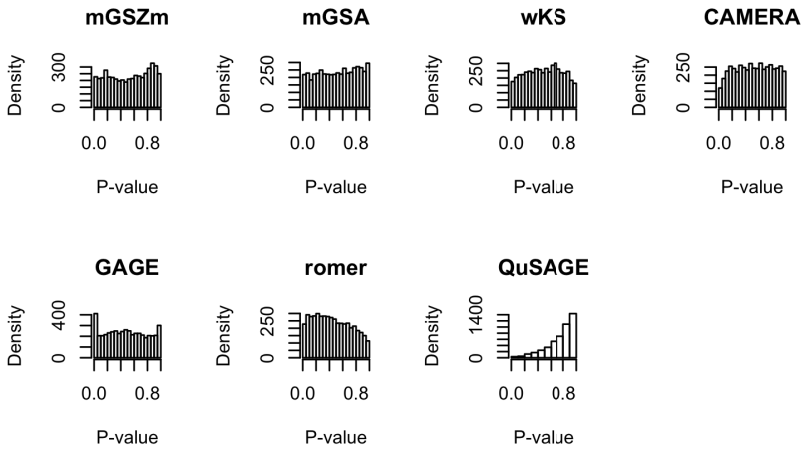


Figure 4.9: Distribution of p-values obtained from mGSZm and six other compared gene set analysis methods with null gene expression data.

Generation of type 1 error

To ensure that mGSZm is not generating false positive results due to the implementation of perm4, we tested mGSZm along with six other gene set analysis methods with null gene expression data. The null gene expression data was generated by randomizing the sample labels of breast cancer gene expression data. Allez was excluded from the evaluation as it does not report P-values. P-values for QuSASE were calculated using the pdf.pVal function in the R/Bioconductor package, *qusage*. As there is no true differential gene expression in null data, p-values generated by an ideal method should follow a uniform distribution. P-values generated by mGSZm followed an approximately uniform distribution, suggesting that mGSZm is fairly conservative in the generation of type 1 errors (Figure 4.9). The P-value distributions of most other methods showed similar distributions except QuSAGE, which is heavily skewed to the left.

4.3 Gene Set Analysis of genome-wide methylation data

Currently, gene set analysis of genome-wide DNA methylation data is predominantly based on over-representation analysis [53, 129, 92, 23] despite the limitations of the approach related to an ambiguous threshold for se-

lecting the list of interesting CpG sites and differential representation of CpGs in genes [39]. Study III of this thesis was based on gene set analysis of genome-wide DNA methylation data with *mGSZ*, a modern signal and ranked list based second generation method developed in study I [82]. *mGSZ* is threshold-free method and thus addresses the problem related to ambiguous thresholds by utilizing whole genome-wide methylation data at the gene level. Methylation levels of CpGs were summarized to gene level by using as proxy the methylation level of CpG within the genomic region of a gene that had maximum absolute t-statistic in differential methylation analysis between active smokers and non-smokers (Section 3.3.4). The novel gene set analysis in this study identified smoking related methylation changes in several novel biological pathways described below.

Among eight major gene set collections in MSigDB, significant results ($FDR \leq 0.05$) were obtained from three - Hallmark, Curated and Gene Ontology (Table 4.4). Results are presented as alterations in methylation patterns in genomic regions associated with biological pathways. Details on hypo- or hyper-methylation is not discussed as the analysis is based on CpGs mapping to any genomic region such as coding or promoter.

4.3.1 GO based gene sets

Eight gene ontology based gene sets were identified to have altered methylation due to smoking. The sense of smell is regulated by several biological processes such as tobacco-induced sinonasal inflammation and squamous cell metaplasia [95]. Unlike other neural tissues, the olfactory epithelium undergoes constant renewal through exfoliation of aged cells and generation of new ones from stem cells [21]. The division and differentiation of the stem cells are regulated by epigenetic mechanisms [43]. The altered methylation pattern (hypermethylation) in genomic regions responsible for olfactory and related G protein coupled receptor activities and pathways identied in this study supports the theory (Tables 4.4). Despite the understanding that smoking is associated with olfactory and gustatory dysfunction, previous studies failed to identify these pathways which could be due to the usage of ORA methods. This study identified a novel biological pathway related to the olfactory system among others in DNA extracted from whole blood, highlighting the importance of the functional class scoring class of GSA methods, which take into account the complete gene list with weights based on the magnitude of the effect.

Smoking has been shown to activate embryonic signalling pathways such as the hedgehog signalling pathway [65]. Smoothened is a protein encoded by SMO gene that plays a role in the hedgehog signalling pathway for cell differentiation. The protein belongs to the class of G protein-coupled receptors. Our results suggest that smoking alters the methylation pattern (hypomethylation) in genes involved in the regulation of the smoothened signalling pathway. We have also identified smoking related altered methylation (hypermethylation) in genes involved in thrombin activated receptor activity. Thrombin receptor activity is related to platelet function and thus plays a role in thrombosis. Active smoking also seems to alter methylation (hypomethylation) in genes involved in semaphorin receptor activity which plays an important role in the developing nervous system [99]. Similarly, genes involved in AP-2 adaptor complex binding are hypermethylated among active smokers of this study. The AP-2 adaptor complex mediates endocytosis which has been shown previously to be affected by smoking [33]. This study pinpoints potential epigenetic mechanism underlining the effect of smoking on endocytosis. This study identified smoking related methylation changes in several G-protein coupled receptor related pathways described above. These are family of proteins that couple with G protein and activate cellular responses to external molecules. Compounds secreted into blood by smoking stimulate the secretion of catecholamine and serotonin. These compounds act on platelets with the help of G protein-coupled receptors, altering the normal functionality of platelets which leads to conditions like thrombosis [73]. Based on the results, we speculate that smoking stimulates G-protein coupled receptor activity through altered methylation (hypermethylation). Similar results were presented by [45].

4.3.2 Curated gene sets

Four curated gene sets were identified to have smoking related alteration in methylation patterns. Cigarette smoke contains compounds that activate the aryl hydrocarbon receptor (AHR) [113]. The activation has been shown to mediate chronic inflammation and cancer [78]. Our results pinpoint a potential epigenetic mechanism involved in AHR activation. Acidic and rich in cysteine (SPARC) is a secreted protein expressed in several cancers. It has been shown that down regulation of SPARC in glioma cells is associated with decreased tumor cell survival [108]. Smoking related alteration in the methylation pattern in SPARC and related genes as identified in this study and shown by others [36] provides insights into the epigenetic basis of smoking effects on cancers. RORA (retinoic acid-related orphan receptor),

a protein that plays an important role in gene regulation, is highly expressed in Chronic obstructive pulmonary disease patients. The gene is known to contribute to emphysema as well as to lung cancer [109]. The finding of methylation changes in genes associated with RORA activation perhaps indicates epigenetic mechanism of smoking related changes in the pathway. Isomers of retinoid receptors - retinoic acid receptor (RAR) and retinoid X receptor (RXR) - are transcription factors important for lung tissue maintanence and repair. The receptors have inhibitory effects on non-small cell lung cancer cell growth [16]. A study has shown that smoking diminishes retinoid acid signalling in lungs [123]. Based on this study, genes related to RARRXR pathway are hypermethylated among smokers, supporting the hypothesis that this could down-regulate the pathway activity as shown by other studies.

4.3.3 Hallmark gene sets

The gene set "UV response down" contains genes that are down-regulated by ultraviolet radiation. Ultraviolet radiation has been previously shown to affect DNA methylation [6]. Interestingly, active smoking seems to affect genes that are sensitive to ultraviolet radiation and this is the first report of such a finding in our knowledge which suggests that active smoking and ultraviolet radiation activate a common pathway.

4.3.4 Results adjusted for alcohol usage and socioeconomic status

Alcohol usage [124] and socioeconomic status [79] have been shown previously to have an effect on DNA methylation. This section presents results of active smoking on DNA methylation adjusted for alcohol usage and socioeconomic status in addition to the covariates mentioned in Section 3.3.4. However, 16 out of 125 participants did not have information on alcohol usage and socioeconomic status. The number of active smokers in the analysis was reduced from 21 to 16 because of missing information on socioeconomic status. Alcohol consumption was measured by asking participants to report their alcohol consumption during the previous week. One unit is equivalent to 14 g of alcohol [55]. Socioeconomic status was based on occupation and was categorized as manual, lower non-manual and upper non-manual. As reduction in sample size due to the additional covariates is likely to reduce the statistical power, we chose to report findings with FDR$\leq$ 0.25 as significant. Note that FDR$\leq$ 0.25 is a commonly used threshold for significant finding in exploratory analyses such as gene set analysis.

Table 4.4: List of gene sets from Molecular Signature Database identified in this study with $FDR \leq 0.05$.

Number of findings	Significant gene sets	Number of genes	GSZ-score	P-value	FDR	Related previous findings
Gene Ontology (GO) based gene sets						
1.	Regulation of smoothened signaling pathway involved in dorsal ventral neural tube (biological process)	8	4.89	5.12e-06	0.04	Related to G-protein coupled receptor [45].
2.	Thrombin activated receptor activity (molecular function)	5	4.27	2.3e-05	0.03	Related to G-protein coupled receptor [52].
3.	Protein domain specific binding (molecular function)	657	4.75	5.7e-05	0.03	[64]
4.	Olfactory receptor activity (molecular function)	321	9.96	6.09e-05	0.03	**Novel finding**
5.	Semaphorin receptor activity (molecular function)	9	4	6.85e-05	0.03	Involved in axonal guidance [116]
6.	AP-2 adaptor complex binding (molecular function)	7	4.22	1.15e-04	0.03	Active smoking affects endocytosis [33].
7.	Nuclear receptor activity (molecular function)	45	4.81	1.18e-04	0.03	[91], [101]
8.	G-protein coupled receptor activity (molecular function)	725	7.21	2.28e-04	0.05	[53]
Curated gene sets						
9.	Aryl hydrocarbon receptor signalling (canonical reactome)	7	5.76	6.27e-07	0.003	[101]
10.	SHI SPARC targets up (chemical and genetic perturbations)	22	5.86	1.22e-06	0.003	[36]
11.	RORA activates gene expression (canonical reactome)	17	4.99	1.68e-05	0.03	[109]
12.	RARRXR pathway (canonical, biocarta)	7	4.38	3.41e-05	0.05	[102]
Hallmark gene sets						
13.	UV response down	138	4.50	0.001	0.05	**Novel finding**

The novel finding of this study - *smoking related altered methylation in olfactory system* - was retained also after adjusting for alcohol usage and socioeconomic status of the participants with FDR ≤ 0.25 (Table 4.5). The well established effect of smoking on aryl hydrocarbon receptor signalling is also retained with FDR of 0.004. Similarly, the effect of smoking on UV sensitive genes as well as four other pathways with gene regulatory effect were retained (Table 4.5).

4.3.5 Strengths, limitations and future direction of the study

To our knowledge, this is the first study to perform second generation gene set analysis of genome-wide methylation data. The approach identified several novel biological pathways with epigenetic signatures induced by active smoking. A major limitation of the study was the sample size, due to which the significant findings were limited to only 13 biological pathways. An ideal sample size for novel findings with a novel approach like this would be similar to the study by [53] that involved meta-analysis of genome-wide

Table 4.5: List of gene sets from Molecular Signature Database identified with additional adjustment with alcohol usage and socioeconomic status with $FDR \leq 0.25$.

Number of findings	**Gene sets**	**FDR**
	Gene ontology (GO) based gene sets	
1.	Thrombin activated receptor activity	0.07
2.	G protein coupled receptor activity	0.08
3.	Olfactory receptor activity	0.14
4.	Protein domain specific binding	0.22
	Curated gene sets	
5.	Aryl hydrocarbon receptor signalling	0.004
6.	SHI SPARC targets up	0.16
	Hallmark gene sets	
7.	UV response down	0.08

DNA methylation measured on 15907 blood-derived DNA samples from participants in 16 cohorts. An integrative analysis of genome-wide DNA methylation and transcriptomics data on the same participants is essential for understanding smoking induced alterations in DNA methylation and its consequences on the transcriptome. An integrative analysis could be done at gene level by associating methylation levels in different genomic regions of a gene and the transcriptomic profile of the same gene. Gene set analysis can improve the power of such a study by analyzing the associations at biological pathway level. For example, separate gene set analyses for CpGs mapping to separate genomic regions such as coding or promoter regions can provide a list of pathways that are hypo- or hypermethylated in specific genomic regions of genes related to a pathway among smokers. Similarly, gene set analysis of transcriptomic data of the same participants can provide a list of pathways up or down regulated among smokers. Analyzing the methylation pattern (hypo- or hypermethylation) in a pathway and the direction of differential expression of the same pathway can provide insights into epigenetic mechanisms by which smoking affects biological pathways.

Chapter 5

Summary and conclusions

The present study was performed to develop an efficient p-value estimation method for permutation based GSA and an advanced permutation method for sample permutation based GSA of complex experimental data with more than two sample groups but less than six replicates per group. Further, the GSA method developed in study I was used to identify smoking related epigenetic signatures in the YFS cohort.

The main achievements of this study are:

I. Publication I addressed the problem of p-value estimation with sample permutation. The published method requires considerably fewer permutations (~500) to accurately estimate p-values and thus speeds up the gene set analysis process. In addition, this study presents novel and robust ways to evaluate GSA methods.

II. Publication II addressed the problem of sample permutation in multi-group data with fewer than six replicates. The study introduced an advanced permutation method designed for data in question and presented an improved evaluation method based on data splitting.

III. Publication III covered the application of the gene set analysis method introduced in publication I to identify alterations in genome-wide methylation patterns due to active smoking at the biological pathway level. The flagship finding of the study is that the highly regenerating olfactory sensing system responds to tobacco smoke and toxin exposure through epigenetic mechanisms.

Overall, the GSA approach holds great potential in the analysis of high-throughput biological datasets as it captures the complex nature of biology where molecules work together. This thesis contributed an improvement of permutation based GSA methods and also introduced an exemplary implementation of a state-of-the-art GSA method for the analysis of genome-wide methylation data with novel discoveries. Nonetheless, several limitations still exist in GSA which need to be addressed.

Chapter 6

Future perspectives

The popularity and success of GSA in bioscience research across several omics platform have motivated the development of more robust methods as well as improvement of existing methods over [114, 34, 75, 127, 96, 67]. The GSA method has led to numerous biological discoveries. For example, GSA of colorectal cancer datasets and glioblastoma multiforme datasets by Drier et al. led to the identification of several gene sets significantly associated with survival of glioblastoma patients and two gene sets with predictive ability for survival time in colorectal cancer [31]. Lee et al. identified mutations responsible for metastatic breast cancers with a GSA based approach [63]. Recently, Shen et al. used a GSA approach to develop immune-based prognostic signatures to predict survival time in ovarian cancer patients [107]. Similarly, with GSA, Gaspar et al. identified several drug classes associated with major depressive disorder with drug repurposing potential [37]. Thus, the implications of GSA range from quick and dirty hypothesis generating exploratory analysis to the identification of molecular signatures for diseases with predictive and prognostic value as well as drug discovery.

Study III of this thesis provided proof of concept that state-of-the-art second generation gene set analysis methods are more powerful for the analysis of epigenome-wide methylation data as compared to the predominantly used overrepresentation based gene set analysis. HumanMethylation450 BeadChip has been replaced by Illumina, Inc. with the HumanMethylationEPIC BeadChip (EPIC), which is based on the same technology but has increased genome coverage measuring methylation at >850,000 CpG sites. GSA of methylation data generated with the improved array is more reliable and reproducible which is crucial in understanding epigenetic mechanisms in health and disease [93].

Next generation sequencing (NGS) technologies allow sequencing of the complete set of miRNAs as well as mRNAs present in an RNA sample. With the NGS driven increase in the amount of miRNAomics data, there is growing demand of statistical methods including GSA methods. GSA based integrative analysis of such data can be used to study interactome and affected pathways of a disease that could potentially point out candidate markers for miRNA-based therapies [27]. GSA of miRNAomics has several limitations at present, such as; i) most of such studies are based on ORA [66, 27], ii) analysis is done indirectly with target mRNAs which could lead to inaccurate results [40], and iii) FCS based GSA methods developed for miRNAomics are based on gene permutation based GSEA which can potentially lead to a high false positive rate [7]. Thus, there is clear need for improvements in GSA methods as well as miRNA categories tailored for miRNAomics data analysis.

The identification of molecular biomarkers that can help in prediction, diagnosis, or treatment of disease is an active field of research in genomic medicine. The conventional method of identifyimg biomarkers from bulk omics data from pooled cell populations can only provide answers at a general level and misses cell specific information. Single-cell RNA sequencing (scRNA-seq) technologies allow the study of gene expression at single-cell resolution revealing the cellular heterogeneity of complex biological systems underlying a disease. This allows the identification of marker genes specific to organs [38]. However, rapid development of this field faces data analysis related challenges as scRNA-seq data are noisier and complex as compared to bulk RNA-seq data [19]. While more FCS based robust GSA methods tailored for scRNA-seq data are needed [74], the methods will help to elucidate several crucial questions such as biological process activities across different cell types [29].

There are several other annotation as well as methodological limitations of GSA [57]. The current knowledge bases for gene sets are based on low-resolution information limited to whether or not a gene is active in a biological pathway. Information at transcripts and single nucleotide variants level generated by large numbers of RNA-seq and genome-wide association analyses are missing in the current knowledge bases. One of the major issues with current GSA is the lack of benchmark datasets for comparing different methods. While development of new methods or improvement of existing methods is important, the presence of several methods in the literature creates confusion among end users in choosing one for their particular data

and interest. Furthermore, as the performance of a method can be data and interest specific, there is no standard way to compare or evaluate GSA methods. This problem has been addressed by several workarounds such as evaluating the methods with: i) artificial data [127, 128], ii) real biological data using a priori knowledge from literature [114], iii) real biological data on the same experiment done at multiple research centers [114, 128], and iv) concordance test between the results from test and training sets from a real biological data [34]. This thesis proposes that multiple evaluation tests should be used by developers for evaluation and that the overall result should be considered as the basis for ranking of methods. This approach minimizes end-users' confusion by presenting evaluation results from different perspectives/methods.

Acknowledgements

This thesis was conducted in the Department of Biosciences at the University of Helsinki, during the years 2012-2016 and in the Department of Clinical Chemistry and Fimlab laboratories at the Tampere University, Faculty of Medicine and Health Technology, during the years 2017-2019.

First and foremost, I would like to express my utmost gratitude to my supervisors Professor Liisa Holm, Dr. Petri Törönen and Professor Terho Lehtimäki for their invaluable support and guidance throughout my study period. Thank you Liisa for providing me the PhD position and for giving me the freedom of thought around my study topics. Thank you Petri for your numerous teaching sessions and inspiring conversations over the years. I want to thank Terho for providing me the idea and opportunity to implement Gene Set Analysis methods developed during the study to analyze genome-wide methylation data from YFS subjects. My grateful acknowledgements to the reviewers of this thesis, Professor Sangita Kulathinal and Assistant Professor Juulia Jylhävä, for their constructive comments of the manuscript.

My thesis committee members, Professor Juho Rousu and Assistant Professor Pekka Marttinen, are warmly thanked for their help and advice during the study follow-up sessions.

I would like to thank all my co-authors for their support and advice. I thank Dr. Alan Medlar for his help in improving my scientific writing. Many thanks to Adjunct Professor Emma Raitaharju and Dr. Nina Mononen for helping me understand the basics of epigenetics and for familiarizing me with YFS data. I also thank Dr. Ismo Hänninen for providing biological insights into epigenetic signatures of active smoking identified with Gene Set Analysis. I would also like to acknowledge Professor Olli Raitakari, Professor Mika Kähönen and Professor Mikko Hurme.

This thesis was funded by Institute of Biotechnology, University of Helsinki; the Academy of Finland; EU Horizon 2020 program (grant 755320 for TAXINOMISIS).

The Young Finns Study has been financially supported by the Academy of Finland: grants 322098, 286284, 134309 (Eye), 126925, 121584, 124282, 129378 (Salve), 117787 (Gendi), and 41071 (Skidi); the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility area of Kuopio, Tampere and Turku University Hospitals (grant X51001); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research ; Finnish Cultural Foundation; The Sigrid Juselius Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahnsson Foundation; Signe and Ane Gyllenberg Foundation; Diabetes Research Foundation of Finnish Diabetes Association; EU Horizon 2020 (grant 755320 for TAXINOMISIS and grant 848146 for TO AITION); European Research Council (grant 742927 for MULTIEPIGEN project); and Tampere University Hospital Supporting Foundation.

References

[1] O Alsheich-Bartok, S Haupt, I Alkalay-Snir, S Saito, E Appella, and Y Haupt. Pml enhances the regulation of p53 by ck1 in response to dna damage. *Oncogene*, 27(26):3653–3661, 2008.

[2] Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance, 2019.

[3] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature genetics*, 30(1):41, 2001.

[4] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[5] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[6] Stella Aslibekyan, Hassan S Dashti, Toshiko Tanaka, Jin Sha, Luigi Ferrucci, Degui Zhi, Stefania Bandinelli, Ingrid B Borecki, Devin M Absher, Donna K Arnett, et al. Prkcz methylation is associated with sunlight exposure in a north american but not a mediterranean population. *Chronobiology international*, 31(9):1034–1040, 2014.

[7] Christina Backes, Qurratulain T Khaleeq, Eckart Meese, and Andreas Keller. mieaa: microrna enrichment analysis and annotation. *Nucleic acids research*, 44(W1):W110–W116, 2016.

[8] Richard T Barfield, Varun Kilaru, Alicia K Smith, and Karen N Conneely. Cpgassoc: an r function for analysis of dna methylation microarray data. *Bioinformatics*, 28(9):1280–1281, 2012.

[9] Emma L B Barrett and David S Richardson. Sex differences in telomeres and lifespan. *Aging cell*, 10(6):913–21, 2011.

[10] William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–1949, 2005.

[11] Tim Beißbarth and Terence P Speed. Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[12] Yoram Ben-Shaul, Hagai Bergman, and Hermona Soreq. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, 21(7):1129–1137, 2005.

[13] Marina Bibikova, Bret Barnes, Chan Tsan, Vincent Ho, Brandy Klotzle, Jennie M Le, David Delano, Lu Zhang, Gary P Schroth, Kevin L Gunderson, et al. High density dna methylation array with single cpg site resolution. *Genomics*, 98(4):288–295, 2011.

[14] Marina Bibikova, Jennie Le, Bret Barnes, Shadi Saedinia-Melnyk, Lixin Zhou, Richard Shen, and Kevin L Gunderson. Genome-wide dna methylation profiling using infinium® assay. *Epigenomics*, 1(1):177–200, 2009.

[15] Marina Bibikova, Zhenwu Lin, Lixin Zhou, Eugene Chudin, Eliza Wickham Garcia, Bonnie Wu, Dennis Doucet, Neal J Thomas, Yunhua Wang, Ekkehard Vollmer, et al. High-throughput dna methylation profiling using universal bead arrays. *Genome research*, 16(3):383–393, 2006.

[16] Jan Brabender, Ralf Metzger, Dennis Salonga, Kathleen D Danenberg, Peter V Danenberg, Arnulf H Hölscher, and Paul M Schneider. Comprehensive expression analysis of retinoic acid receptors and retinoid x receptors in non-small cell lung cancer: implications for tumor development and prognosis. *Carcinogenesis*, 26(3):525–530, 2005.

[17] Rainer Breitling, Anna Amtmann, and Pawel Herzyk. Iterative group analysis (iga): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC bioinformatics*, 5(1):34, 2004.

[18] Sebastien Cagnol and Jean-Claude Chambard. Erk and cell death: mechanisms of erk-induced cell death–apoptosis, autophagy and senescence. *The FEBS journal*, 277(1):2–21, 2010.

[19] Geng Chen and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, 10:317, 2019.

[20] C L Cheney, P Lenssen, S N Aker, B A Cunningham, J M Gauvreau, J Darbinian, and K V Barale. Sex differences in nitrogen balance following marrow grafting for leukemia. *Journal of the American College of Nutrition*, 6(3):223–230, 1987.

[21] Rhea Choi and Bradley J Goldstein. Olfactory epithelium: cells, clinical disorders, and insights from an adult stem cell niche. *Laryngoscope investigative otolaryngology*, 3(1):35–42, 2018.

[22] Francis S Collins and Victor A McKusick. Implications of the human genome project for medical science. *Jama*, 285(5):540–544, 2001.

[23] Ashley L Comes, Darina Czamara, Kristina Adorjan, Heike Anderson-Schmidt, Till FM Andlauer, Monika Budde, Katrin Gade, Maria Hake, Janos L Kalman, Sergi Papiol, et al. The role of environmental stress and dna methylation in the longitudinal course of bipolar disorder. *International journal of bipolar disorders*, 8(1):1–12, 2020.

[24] Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic acids research*, 45(D1):D331–D338, 2016.

[25] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[26] G M Cooper and G Nayak. p53 is a major component of the transcriptional and apoptotic program regulated by pi 3-kinase/akt/gsk3 signaling. *Cell Death and Disease*, 3(10):e400, 2012.

[27] Rodrigo Coutinho de Almeida, Yolande FM Ramos, AMETA Mahfouz, Wouter den Hollander, Nico Lakenberg, Evelyn Houtman, Marcella van Hoolwerff, H Eka D Suchiman, Alejandro Rodriguez-Ruiz, P Eline Slagboom, et al. Rna sequencing data integration reveals an mirna interactome of osteoarthritis cartilage. *Annals of the Rheumatic Diseases: an international peer-reviewed journal for health professionals and researchers in the rheumatic diseases*, 2018.

[28] David Croft, Gavin O'kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697, 2010.

[29] Hongxu Ding, Andrew Blair, Ying Yang, and Joshua M Stuart. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nature communications*, 10(1):1–6, 2019.

[30] Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Improving gene set analysis of microarray data by sam-gs. *BMC bioinformatics*, 8(1):242, 2007.

[31] Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

[32] Lina Du, HÃ1lya Bayir, Yichen Lai, Xiaopeng Zhang, Patrick M Kochanek, Simon C Watkins, Steven H Graham, and Robert S B Clark. Innate gender-based proclivity in response to cytotoxicity and programmed cell death pathway. *The Journal of biological chemistry*, 279(37):38563–38570, 2004.

[33] P Duffney, TH Thatcher, RP Phipps, and PJ Sime. Cigarette smoke alters endocytosis in primary human small airway epithelial cells. In *B61. EPITHELIAL CELL BIOLOGY IN RESPIRATORY DISEASE*, pages A3809–A3809. American Thoracic Society, 2018.

[34] Bradley Efron, Robert Tibshirani, et al. On testing the significance of sets of genes. *The annals of applied statistics*, 1(1):107–129, 2007.

[35] Zhaohui Feng. p53 regulation of the igf-1/akt/mtor pathways and the endosomal compartment. *Cold Spring Harbor perspectives in biology*, 2(2):a001057, 2010.

[36] Jun Gao, Jian Song, Haojie Huang, Zhaoshen Li, Yiqi Du, Jia Cao, Minghui Li, Shunli Lv, Han Lin, and Yanfang Gong. Methylation of the sparc gene promoter and its clinical implication in pancreatic cancer. *Journal of Experimental & Clinical Cancer Research*, 29(1):28, 2010.

[37] Héléna A Gaspar, Zachary Gerring, Christopher Hübel, Christel M Middeldorp, Eske M Derks, and Gerome Breen. Using genetic drug-target networks to develop new drug hypotheses for major depressive disorder. *Translational psychiatry*, 9(1):117, 2019.

[38] Danuta R Gawel, Jordi Serra-Musach, Sandra Lilja, Jesper Aagesen, Alex Arenas, Bengt Asking, Malin Bengnér, Janne Björkander, Sophie Biggs, Jan Ernerudh, et al. A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. *Genome medicine*, 11(1):47, 2019.

[39] Paul Geeleher, Lori Hartnett, Laurance J Egan, Aaron Golden, Raja Affendi Raja Ali, and Cathal Seoighe. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics*, 29(15):1851–1857, 2013.

[40] Patrice Godard and Jonathan van Eyll. Pathway analysis from lists of microrna: common pitfalls and alternative strategy. *Nucleic acids research*, 43(7):3490–3497, 2015.

[41] Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[42] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[43] Bradley J Goldstein, Garrett M Goss, Rhea Choi, Dieter Saur, Barbara Seidler, Joshua M Hare, and Nirupa Chaudhari. Contribution of polycomb group proteins to olfactory basal stem cell self-renewal in a novel c-kit+ culture model and in vivo. *Development*, 143(23):4394–4404, 2016.

[44] Maxim VC Greenberg and Deborah Bourcâhis. The diverse roles of dna methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, pages 1–18, 2019.

[45] Tina Haase, Christian Müller, Julia Krause, Caroline Röthemeier, Justus Stenzig, Sonja Kunze, Melanie Waldenberger, Thomas Münzel, Norbert Pfeiffer, Philipp Wild, et al. Novel dna methylation sites influence gpr15 expression in relation to smoking. *Biomolecules*, 8(3):74, 2018.

[46] annotation Hansen K. ilmn12. hg19: annotation for illumina's 450k methylation arrays. 2015. *R package, version 0.2*, 2015.

[47] L Alexis Hoeferlin, Baharan Fekry, Besim Ogretmen, Sergey a Krupenko, and Natalia I Krupenko. Folate stress induces apoptosis via p53-dependent de novo ceramide synthesis and up-regulation of ceramide synthase 6. *The Journal of biological chemistry*, 2013.

[48] Sonja Hombach and Markus Kretz. Non-coding rnas: classification, biology and functioning. In *Non-coding RNAs in Colorectal Cancer*, pages 3–17. Springer, 2016.

[49] Eugene Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Carmen J Marsit, Heather H Nelson, John K Wiencke, and Karl T Kelsey. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics*, 13(1):86, 2012.

[50] Da Wei Huang, Brad T Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W Baseler, H Clifford Lane, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(suppl_2):W169–W175, 2007.

[51] Rafael A Irizarry, Chi Wang, Yun Zhou, and Terence P Speed. Gene set enrichment analysis made simple. *Statistical methods in medical research*, 18(6):565–575, 2009.

[52] Min A Jhun, Jennifer A Smith, Erin B Ware, Sharon LR Kardia, Thomas H Mosley Jr, Stephen T Turner, Patricia A Peyser, and Sung Kyun Park. Modeling the causal role of dna methylation in the association between cigarette smoking and inflammation in african americans: a 2-step epigenetic mendelian randomization study. *American journal of epidemiology*, 186(10):1149–1158, 2017.

[53] Roby Joehanes, Allan C Just, Riccardo E Marioni, Luke C Pilling, Lindsay M Reynolds, Pooja R Mandaviya, Weihua Guan, Tao Xu,

Cathy E Elks, Stella Aslibekyan, et al. Epigenetic signatures of cigarette smoking. *Circulation: Cardiovascular Genetics*, 9(5):436–447, 2016.

[54] Peter A Jones. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484, 2012.

[55] Markus Juonala, Jorma SA Viikari, Mika Kähönen, Tomi Laitinen, Leena Taittonen, Britt-Marie Loo, Antti Jula, Jukka Marniemi, Leena Räsänen, Tapani Rönnemaa, et al. Alcohol consumption is directly associated with carotid intima–media thickness in finnish young adults: The cardiovascular risk in young finns study. *Atherosclerosis*, 204(2):e93–e98, 2009.

[56] Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[57] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375, 2012.

[58] Seon-Young Kim and David J Volsky. Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, 6(1):144, 2005.

[59] Theo A Knijnenburg, Lodewyk FA Wessels, Marcel JT Reinders, and Ilya Shmulevich. Fewer permutations, more accurate p-values. *Bioinformatics*, 25(12):i161–i168, 2009.

[60] Sek Won Kong, William T Pu, and Peter J Park. A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, 22(19):2373–2380, 2006.

[61] Samuel Kotz and Saralees Nadarajah. *Extreme value distributions: theory and applications*. World Scientific, 2000.

[62] B J Kroesen, B Pettus, C Luberto, M Busman, H Sietsma, L de Leij, and Y A Hannun. Induction of apoptosis through b-cell receptor cross-linking occurs via de novo generated c16-ceramide and involves mitochondria. *The Journal of biological chemistry*, 276(17):13606–13614, 2001.

[63] Ji-Hyun Lee, Xing-Ming Zhao, Ina Yoon, Jin Young Lee, Nam Hoon Kwon, Yin-Ying Wang, Kyung-Min Lee, Min-Joo Lee, Jisun Kim, Hyeong-Gon Moon, et al. Integrative analysis of mutational and

transcriptional profiles reveals driver mutations of metastatic breast cancers. *Cell discovery*, 2:16025, 2016.

[64] Ken WK Lee and Zdenka Pausova. Cigarette smoking and dna methylation. *Frontiers in genetics*, 4:132, 2013.

[65] Hassan Lemjabbar-Alaoui, Vijay Dasari, Sukhvinder S Sidhu, Aklilu Mengistab, Walter Finkbeiner, Marianne Gallup, and Carol Basbaum. Wnt and hedgehog are critical mediators of cigarette smoke-induced lung cancer. *PloS one*, 1(1):e93, 2006.

[66] Wentong Li, Yalan Yang, Ying Liu, Shuai Liu, Xiuxiu Li, Yingping Wang, Yanmin Zhang, Hui Tang, Rong Zhou, and Kui Li. Integrated analysis of mrna and mirna expression profiles in livers of yimeng black pigs with extreme phenotypes for backfat thickness. *Oncotarget*, 8(70):114787, 2017.

[67] Yiqun Li, Ying Wu, Xiaohan Zhang, Yunfan Bai, Luqman Muhammad Akthar, Xin Lu, Ming Shi, Jianxiang Zhao, Qinghua Jiang, and Yu Li. Scia: A novel gene set analysis applicable to data with different characteristics. *Frontiers in Genetics*, 10, 2019.

[68] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[69] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[70] Fudong Liu, Zhong Li, Jun Li, Chad Siegel, Rongwen Yuan, and Louise D McCullough. Sex differences in caspase activation after stroke. *Stroke; a journal of cerebral circulation*, 40(5):1842–1848, 2009.

[71] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3):535–550, 2016.

[72] David J Lockhart, Helin Dong, Michael C Byrne, Maximillian T Follettie, Michael V Gallo, Mark S Chee, Michael Mittmann, Chunwei

Wang, Michiko Kobayashi, Heidi Norton, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675, 1996.

[73] Curtis Lee Lowery III, Clay Elliott, Anthonya Cooper, Coedy Hadden, Roberto N Sonon, Parastoo Azadi, D Keith Williams, James D Marsh, Donna S Woulfe, and Fusun Kilic. Cigarette smoking-associated alterations in serotonin/adrenalin signaling pathways of platelets. *Journal of the American Heart Association*, 6(5):e005465, 2017.

[74] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6), 2019.

[75] Weijun Luo, Michael S Friedman, Kerby Shedden, Kurt D Hankenson, and Peter J Woolf. Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):161, 2009.

[76] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mrna and protein in complex biological samples. *FEBS letters*, 583(24):3966–3973, 2009.

[77] Ulrich Mansmann and R Meister. Testing differential gene expression in functional groups. *Methods of information in medicine*, 44(03):449–453, 2005.

[78] CA Martey, CJ Baglole, TA Gasiewicz, PJ Sime, and RP Phipps. The aryl hydrocarbon receptor is a regulator of cigarette smoke induction of the cyclooxygenase and prostaglandin pathways in human lung fibroblasts. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 289(3):L391–L399, 2005.

[79] Thomas W McDade, Calen P Ryan, Meaghan J Jones, Morgan K Hoke, Judith Borja, Gregory E Miller, Christopher W Kuzawa, and Michael S Kobor. Genome-wide analysis of dna methylation in relation to socioeconomic status during development and early adulthood. *American journal of physical anthropology*, 169(1):3–11, 2019.

[80] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, et al. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, 2015.

[81] Jorge R Mera, Beverly Dickson, and Mark Feldman. Influence of gender on the ratio of serum aspartate aminotransferase (ast) to alanine aminotransferase (alt) in patients with and without hyperbilirubinemia. *Digestive diseases and sciences*, 53(3):799–802, 2008.

[82] Pashupati Mishra, Petri Törönen, Yrjö Leino, and Liisa Holm. Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics*, 30(19):2747–2756, 2014.

[83] Cristina Mitrea, Zeinab Taghavi, Behzad Bokanizad, Samer Hanoudi, Rebecca Tagett, Michele Donato, Calin Voichita, and Sorin Draghici. Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in physiology*, 4:278, 2013.

[84] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267, 2003.

[85] Antonio Mora. Gene set analysis methods for the functional interpretation of non-mrna dataâgenomic range and ncrna data. *Briefings in bioinformatics*, 2019.

[86] Kai-Oliver Mutz, Alexandra Heilkenbrinker, Maren Lönne, Johanna-Gabriela Walter, and Frank Stahl. Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, 24(1):22–30, 2013.

[87] Haroon Naeem, Ralf Zimmer, Pegah Tavakkolkhah, and Robert KÃ¼ffner. Rigorous assessment of gene set enrichment tests. *Bioinformatics*, 28(11):1–7, 2012.

[88] Michael A Newton, Fernando A Quintana, Johan A Den Boon, Srikumar Sengupta, and Paul Ahlquist. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Annals of Applied Statistics*, 1(1):85–106, 2007.

[89] Cornelia O'Callaghan-Sunol, Vladimir L Gabai, and Michael Y Sherman. Hsp27 modulates p53 signaling and suppresses cellular senescence. *Cancer research*, 67(24):11779–11788, 2007.

[90] Paul Pavlidis, Darrin P Lewis, and William Stafford Noble. Exploring gene expression data with class scores. In *Biocomputing 2002*, pages 474–485. World Scientific, 2001.

[91] Robert A Philibert, SRH Beach, and Gene H Brody. The dna methylation signature of smoking: an archetype for the identification of biomarkers for behavioral illness. In *Genes and the motivation to use substances*, pages 109–127. Springer, 2014.

[92] Rachael V Phillips, Linda Rieswijk, Alan E Hubbard, Roel Vermeulen, Jinming Zhang, Wei Hu, Laiyu Li, Bryan A Bassig, Jason YY Wong, Boris Reiss, et al. Human exposure to trichloroethylene is associated with increased variability of blood dna methylation that is enriched in genes and pathways related to autoimmune disease and cancer. *Epigenetics*, 14(11):1112–1124, 2019.

[93] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, 17(1):208, 2016.

[94] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature biotechnology*, 28(10):1057, 2010.

[95] Edith Puchelle, Jean-Marie Zahm, Jean-Marie Tournier, and Christelle Coraux. Airway epithelial repair, regeneration, and remodeling after injury in chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, 3(8):726–733, 2006.

[96] Wenyi Qin, Xujun Wang, and Hui Lu. A novel joint gene set analysis framework improves identification of enriched pathways in cross disease transcriptomic analysis. *Frontiers in Genetics*, 10:293, 2019.

[97] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[98] Olli T Raitakari, Markus Juonala, Tapani Rönnemaa, Liisa Keltikangas-Järvinen, Leena Räsänen, Matti Pietikäinen, Nina Hutri-Kähönen, Leena Taittonen, Eero Jokinen, Jukka Marniemi, et al. Cohort profile: the cardiovascular risk in young finns study. *International journal of epidemiology*, 37(6):1220–1226, 2008.

[99] Jonathan A Raper. Semaphorins and their receptors in vertebrates and invertebrates. *Current opinion in neurobiology*, 10(1):88–94, 2000.

[100] Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, 14(2):482–517, 2019.

[101] Lindsay M Reynolds, Ma Wan, Jingzhong Ding, Jackson R Taylor, Kurt Lohman, Dan Su, Brian D Bennett, Devin K Porter, Ryan Gimple, Gary S Pittman, et al. Dna methylation of the aryl hydrocarbon receptor repressor associations with cigarette smoking and subclinical atherosclerosis. *Circulation: Cardiovascular Genetics*, 8(5):707–716, 2015.

[102] Mikael V Ringh, Michael Hagemann-Jensen, Maria Needhamsen, Lara Kular, Charles E Breeze, Louise K Sjöholm, Lara Slavec, Susanna Kullberg, Jan Wahlström, Johan Grunewald, et al. Tobacco smoking induces changes in true dna methylation, hydroxymethylation and gene expression in bronchoalveolar lavage cells. *EBioMedicine*, 46:290–304, 2019.

[103] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[104] Tina Rönn, Petr Volkov, Linn Gillberg, Milana Kokosar, Alexander Perfilyev, Anna Louisa Jacobsen, Sine W Jørgensen, Charlotte Brøns, Per-Anders Jansson, Karl-Fredrik Eriksson, et al. Impact of age, bmi and hba1c levels on the genome-wide dna methylation and mrna expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Human molecular genetics*, 24(13):3792–3813, 2015.

[105] Steven L Salzberg. Open questions: How many genes do we have? *BMC biology*, 16(1):94, 2018.

[106] Petr V Sergiev, Anna Ya Golovina, Ilya A Osterman, Michail V Nesterchuk, Olga V Sergeeva, Anastasia A Chugunova, Sergey A Evfratov, Ekaterina S Andreianova, Philipp I Pletnev, Ivan G Laptev,

et al. N6-methylated adenosine in rna: from bacteria to humans. *Journal of molecular biology*, 428(10):2134–2145, 2016.

[107] Sipeng Shen, Guanrong Wang, Ruyang Zhang, Yang Zhao, Hao Yu, Yongyue Wei, and Feng Chen. Development and validation of an immune gene-set based prognostic signature in ovarian cancer. *EBioMedicine*, 40:318–326, 2019.

[108] Q Shi, S Bao, L Song, Q Wu, DD Bigner, AB Hjelmeland, and JN Rich. Targeting sparc expression decreases glioma cellular survival and invasion associated with reduced activities of fak and ilk kinases. *Oncogene*, 26(28):4084, 2007.

[109] Ying Shi, Jiaofei Cao, Jane Gao, Liang Zheng, Andrew Goodwin, Chang Hyoek An, Avignat Patel, Janet S Lee, Steven R Duncan, Naftali Kaminski, et al. Retinoic acid–related orphan receptor-α is induced in the setting of dna damage and promotes pulmonary emphysema. *American journal of respiratory and critical care medicine*, 186(5):412–419, 2012.

[110] Yan Song, Jinsoo Ahn, Yeunsu Suh, Michael E Davis, and Kichoon Lee. Identification of novel tissue-specific genes by analysis of microarray databases: a human and mouse model. *PloS one*, 8(5):e64483, 2013.

[111] You Song, Honglei Ren, and Jinzhi Lei. Collaborations between cpg sites in dna methylation. *International Journal of Modern Physics B*, 31(20):1750243, 2017.

[112] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[113] Russell L Stedman. Chemical composition of tobacco and tobacco smoke. *Chemical Reviews*, 68(2):153–207, 1968.

[114] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[115] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599, 2018.

[116] Hongwei Tang, Peng Wei, Eric J Duell, Harvey A Risch, Sara H Olson, H Bas Bueno-de Mesquita, Steven Gallinger, Elizabeth A Holly, Gloria Petersen, Paige M Bracci, et al. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene-and pathway-based interaction analysis of gwas data. *Carcinogenesis*, 35(5):1039–1045, 2014.

[117] Saeed Tavazoie, Jason D Hughes, Michael J Campbell, Raymond J Cho, and George M Church. Systematic determination of genetic network architecture. *Nature genetics*, 22(3):281, 1999.

[118] Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, and Peter J Park. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, 102(38):13544–13549, 2005.

[119] Petri Törönen, Pauli J Ojala, Pekka Marttinen, and Liisa Holm. Robust extraction of functional signals from gene set analysis using a generalized threshold free scoring function. *BMC Bioinformatics*, 10(1):307, 2009.

[120] Timothy J Triche Jr, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of illumina infinium dna methylation beadarrays. *Nucleic acids research*, 41(7):e90–e90, 2013.

[121] Jose Vina, Juan Gambini, Raul Lopez-Grueso, Khira M Abdelaziz, Mariona Jove, and Consuelo Borras. Females live longer than males: role of oxidative stress. *Current pharmaceutical design*, 17(36):3959–65, 2011.

[122] Kimmo Virtaneva, Fred A Wright, Stephan M Tanner, Bo Yuan, William J Lemon, Michael A Caligiuri, Clara D Bloomfield, Albert De La Chapelle, and Ralf Krahe. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proceedings of the National Academy of Sciences*, 98(3):1124–1129, 2001.

[123] Jianmiao Wang, Wei Liu, Chad Marion, Rajvir Singh, Nathaniel Andrews, Chun Geun Lee, Jack A Elias, and Charles S Dela Cruz. Regulation of retinoic acid receptor beta by interleukin-15 in the lung during cigarette smoking and influenza virus infection. *American journal of respiratory cell and molecular biology*, 53(6):822–833, 2015.

[124] Lauren E Wilson, Zongli Xu, Sophia Harlid, Alexandra J White, Melissa A Troester, Dale P Sandler, and Jack A Taylor. Alcohol and dna methylation: An epigenome-wide association study in blood and normal breast tissue. *American journal of epidemiology*, 188(6):1055–1065, 2019.

[125] Marcia M Wright and Christopher R McMaster. Phospholipid synthesis, diacylglycerol compartmentation, and apoptosis. *Biological research*, 35(2):223–229, 2002.

[126] Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E Visvader, and Gordon K Smyth. Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010.

[127] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

[128] Gur Yaari, Christopher R Bolen, Juilee Thakar, and Steven H Kleinstein. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic acids research*, 41(18):e170–e170, 2013.

[129] Xiaofei Yang, Lin Gao, and Shihua Zhang. Comparative pan-cancer dna methylation analysis reveals cancer common and specific patterns. *Briefings in bioinformatics*, 18(5):761–773, 2017.